Prediction of antibiotic resistance at the patient level using deep learning

Juan S. Inda-Díaz[1,2], Anna Johnning[1,2,3], Magnus Hessel[4], Anders Sjöberg[3,5], Anna Lokrantz[3], Lisa Helldal[4], Mats Jirstrand[3,5], Lennart Svensson[5] and Erik Kristiansson[1,2*]

[1]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg; Gothenburg, SE-41296, Sweden.

[2]Centre for Antibiotic Resistance Research (CARe), University of Gothenburg; Gothenburg, SE-41124, Sweden.

[3]Department of Systems and Data Analysis, Fraunhofer-Chalmers Centre; Gothenburg, SE-41288, Sweden.

[4]Clinical microbiology, Sahlgrenska University Hospital; Gothenburg, SE-41345, Sweden.

[5]Department of Electrical Engineering, Chalmers University of Technology; Gothenburg, SE-41296, Sweden.

*Address correspondence to Erik Kristiansson, erik.kristiansson@chalmers.se.

ABSTRACT

Rapid and accurate diagnostics of bacterial infections are necessary for efficient treatment of antibiotic-resistant pathogens. Cultivation-based methods, such as antibiotic susceptibility testing (AST), are limited by bacterial growth rates and may not produce results before the treatment

starts. This increases patient risks and antibiotic overprescription. Here, we present a deep-learning method that merges patient data with available AST results to predict antibiotic susceptibilities that have not yet been measured. The method is combined with conformal prediction (CP) to enable the estimation of uncertainty of the predictions at the patient level. After training on three million AST results from thirty European countries, the method can predict susceptibility with a major error rate below 2.5% for quinolones, cephalosporins, and aminoglycosides, and below 12% for penicillins. Furthermore, the model predicts resistance of cephalosporins and fluoroquinolones with an average very major rate of 1.5% and 3.2%, respectively, but with higher very major error rates for penicillins, nalidixic acid, and aminoglycosides. We also show that the method reflects empirical error rates, even when limited diagnostic information is available. We conclude that decision support based on deep learning may offer new means to meet the growing burden of antibiotic resistance.

IMPORTANCE

Improved diagnostics tools are vital for maintaining efficient treatments of antibiotic-resistant bacteria and for reducing the overconsumption of antibiotics. In our research, we introduce a new deep learning-based method capable of predicting untested antibiotic resistance phenotypes. The method utilizes transformers – a powerful technique also used in large language models – which can efficiently take advantage of antibiotic susceptibility tests (AST) and patient data simultaneously. The model produces computational predictions that can be used as time- and cost-efficient alternatives to results from additional cultivation-based diagnostic tests. Significantly, our study highlights the potential of AI technologies for meeting the growing burden of antibiotic-resistant bacterial infections.

INTRODUCTION

The global rise of antibiotic-resistant bacterial infections threatens human health globally *(1)*. Today, almost five million yearly deaths are attributed to antibiotic-resistant bacteria *(2)*, a number that is expected to continue to grow in the coming decades *(3)*. Reduced efficiency of antibiotic treatment increases the risk of performing vital healthcare procedures – including surgery, chemotherapy, and organ transplantation *(4)* – and, thereby, jeopardizes modern medicine as a whole.

Accurate and fast diagnostics are necessary for efficient treatment of antibiotic-resistant bacteria. A central method is antibiotic susceptibility testing (AST), a laboratory test in which a bacterium isolated from a patient sample is cultivated and its resistance phenotype assessed in the presence of antibiotics *(5)*. However, AST can be time-consuming due to the often low growth rate of bacteria and the large number of antibiotics that may need to be tested for highly multi-drug resistant isolates. Tiered-based testing strategies and other time constraints make AST results initially incomplete. For life-threatening infections, treatment needs to start as early as possible, often before all test results are available *(6)*. Under these circumstances, the choice of treatment is reduced to educated guesses based on limited diagnostics information *(7)*. This form of "empirical" treatment is associated with increased patient risks and overprescription of antibiotics (8–10).

Antibiotic resistance is commonly caused by resistance genes encoding various defense mechanisms. These genes are often co-localized on mobile genetic elements, such as plasmids and transposons *(11, 12)*. Multiple resistance genes can, thus, be transferred simultaneously between bacterial cells, which gives rise to strong correlations between the susceptibility to different antibiotics. The type of infecting bacterium and, thus, its susceptibility profile, is,

furthermore, dependent on patient characteristics, including age, sex, and the geographical region where the infection was acquired (13–15). Indeed, patient data has previously been shown to contain valuable information for the selection of suitable antibiotic therapy for bacterial urinary tract infections (15–17). There are, however, no methods that can combine patient data with the initial, often incomplete AST results and, thus, guide the choice of treatment based on all available diagnostic information. Indeed, combining patient data with the available AST results could enable more accurate prediction of the susceptibilities to antibiotics that have not been tested and, thereby, provide physicians with comprehensive diagnostic information at an earlier point in time.

Artificial intelligence (AI) and deep learning have been successfully applied to diagnostics *(18, 19)*, but the focus has primarily been on image-based data commonly used in radiology and pathology *(20)*. In contrast, methods for non-image multimodal data, which is dominating in the diagnostics of infectious diseases, have received less attention *(21)*. There are, in fact, currently no AI-based decision support systems for the selection of antibiotic treatment approved by the Food and Drug Administration (https://medicalfuturist.com/fda-approved-ai-based-algorithms/) *(22)*. A major culprit in the development of such methods is the complexity of the diagnostic data, which is typically categorical (stratified test results and patient information) and contains dependencies and redundancies. Also, when applied to antibiotic-resistant bacteria, the incompleteness of AST results requires approaches that can efficiently handle missing data. Furthermore, since the model accuracy depends on the degree of available information, any prediction must be associated with estimates of the prediction uncertainty. Indeed, the possibility of disregarding insufficiently confident predictions is vital in critical decision-making and, thus, essential for the adoption of AI-based methodology in healthcare

settings *(23)*. However, today, most AI-based methods for diagnostics are primarily evaluated on populations and do not provide information about the uncertainty for predictions at the patient level *(24)*. Existing methods have, so far, been unable to properly address these challenges and are either limited to a few specific antibiotics, cannot incorporate missing AST data, or do not provide any estimates of the predictive uncertainty *(25, 26)*.

In recent years, transformer-based models, such as BERT (bidirectional encoder representations from transformers) *(27)* and GPT (generative pre-trained transformer) *(28)*, have transformed natural language processing. These models operate on categorical input data, often structured into sentences of words, and subject them to multi-head self-attention *(29)*. This makes it possible to infer complex dependencies between words directly from data and, thereby, predict parts that are missing. Therefore, we hypothesized that transformers may be suitable for the prediction of antibiotic susceptibility results from a combination of incomplete diagnostic information and patient data. Multimodal self-attention would enable the identification of the complex dependencies between the diagnostic data types and facilitate extrapolation to susceptibilities that have not been tested. Transformers have previously been shown to be highly useful outside natural language processing *(30, 31)*, but are rarely used for diagnostics of infectious diseases.

In this study, we present a novel transformer-based method that can accurately predict antibiotic susceptibilities based on patient data and incomplete diagnostic information. We combine the model with conditional inductive conformal prediction (CICP) *(32)* to estimate the uncertainty of each prediction at the patient level and, hence, abstain from presenting predictions with too low certainty. The model was trained and evaluated on a large heterogeneous dataset containing AST results from blood infections caused by *Escherichia coli* collected from routine

care in thirty European countries. Our results showed that the model could make accurate predictions of susceptibility to a wide range of antibiotics, often when only a few AST results were included in the input. We also show that the model handles increasing information well, providing more accurate predictions as more AST results become available. Finally, we show that predictions can be made within pre-defined confidence levels for most of the bacterial isolates and antibiotics, which allows control of both the major and very major error rates. We conclude that the combination of transformers and CICP constitutes an appealing class of models for the integration and prediction of diagnostic information.

RESULTS

A transformer-based model for antibiotic susceptibility prediction

We developed a transformer-based model to predict unavailable diagnostic information using multiple classifications. The input to the model is a sequence of words containing the available diagnostic information (results of antibiotic susceptibility testing; AST) and patient data (Figure 1). The AST results are assumed to be incomplete, where only a subset of the possible tests have available results. The transformer is used to estimate a sentence-specific classification (*CLS*) vector, which serves as input to multiple antibiotic-specific neural networks, each predicting the probability of susceptibility to the corresponding antibiotic. The uncertainty is estimated through inductive conformal prediction, conditioned on the susceptibility; hence, controlling the false positive and false negative error rates for each antibiotic *(32, 33)*. The full details of the architecture of the model and the uncertainty estimation are provided in Methods.

The model was trained and evaluated on data from The European Surveillance System (TESSy) (https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy), which contains AST results from 468,429 *Escherichia coli* isolates collected from blood infections of patients in thirty European countries. The training and evaluation were done on sixteen commonly used antibiotics that belong to four large and clinically relevant antibiotic classes: aminoglycosides, cephalosporins, penicillins, and quinolones (Table 1). A bacterial isolate was, on average, tested for susceptibility to 7.5 antibiotics with a standard deviation of 1.96 (Figure 2A and B). The most commonly tested antibiotics were ampicillin (AMP), ceftazidime (CAZ), cefotaxime (CTX), ciprofloxacin (CIP), and gentamicin (GEN), for which at least 79% of the isolates were tested, regardless of the country of origin or gender of the patient. In contrast, less than 5% of the bacterial isolates were tested for piperacillin (PIP), moxifloxacin (MFX), nalidixic acid (NAL), and ofloxacin (OFX); indeed, the median number of isolates tested for these antibiotics among all countries did not exceed 0.1% for any of these antibiotics (Figure 2C). The rate of susceptibility for each isolate and country was lowest for the penicillins amoxicillin (AMX), AMP, and PIP (37%-50%). For other antibiotics, the susceptibility rates were more unbalanced: cephalosporins and aminoglycosides had 89%-93% susceptible bacterial isolates. A slightly higher rate of susceptibility was observed for isolates collected from female patients (Figure 2D).

Predictions of antibiotic susceptibility have high performance

To test how well the model predicts the susceptibility of antibiotics that were not included as input to the model, we trained the model on 80% of the bacterial isolates, with 10% reserved exclusively for calibration and 10% for testing. During training, calibration, and testing, we selected a random number of antibiotics as input to the model (distribution available in

Figure 3A, mean 6.03 with a standard deviation of 1.42; see Methods for details) together with all the patient information. The susceptibility of the remaining tested antibiotics was assumed to be unknown and, therefore, removed from the input data. The predictions produced by the model were then compared to the antibiotics removed from the input to evaluate the model performance. In this setting, the model had an overall high performance that did not differ substantially between training and test data (Figure 3B). There were, however, clear differences in performance between antibiotics. The F1 score (the harmonic mean of precision and recall) was highest for cephalosporins (83%-89%, average 86%) and quinolones (73%-89%, average 83%), though lower for penicillins (63%-92%, average 78%) when excluding the combination treatment of piperacillin/tazobactam (TZP), 42%. The aminoglycosides had the lowest F1 scores: 47% for GEN and 61% for tobramycin (TOB).

Next, we evaluated the model based on the major error (ME) and very major error (VME) rates, defined as the proportion of the susceptible and resistant bacterial isolates being erroneously predicted, respectively. The ME and VME rates are common performance measures in antimicrobial susceptibility testing and are frequently used to evaluate and compare diagnostics methods *(34)*. Based on the test dataset, cephalosporins had, on average, an ME rate of 1.7% (0.9%–2.0%) while the VME rate was, on average, 12.0% (6.8%–18.8%) (Figure 3C and D). For quinolones, the average ME rate was 3.5% (2.1%–5.6%), and the VME rate for levofloxacin (LVX), MFX, and OFX, was, on average, 11.5% (9.9%–12.4%), while it was considerably higher for CIP and nalidixic acid (NAL) (31.1% and 32.2%, respectively). The penicillins had an overall higher average ME rate of 13.5% (5.3%-19.4%). Here, PIP and AMX had the lowest VME rates (9.7%, and 17.6%, respectively) while the rest of the penicillins had an average VME rate of 32.8% (30.9%–34.6%). To investigate the effect of the patient data, we

compared our results with a model based only on AST results and no patient data (Figure S1). The full model had an overall lower VME, especially for cephalosporins (average reduction of 45%), aminoglycosides (20%), and quinolones (20%), showing that including the patient data improves the model performance.

The performance of the model was heavily influenced by the number of AST results that were included in the input sentence. When the number of AST results used as input to the model increased from four to eight antibiotics, large reductions were seen in the ME rate for penicillins, where AMX and AMP dropped from 41% and 21% to 5% and 7%, respectively (Figure 3D). Interestingly, the drop for amoxicillin/clavulanic acid (AMC) was more modest (from 17% to 14%), and the ME rate for TZP did not decrease at all. A substantial reduction could also be seen in the VME rate for all antibiotics except AMX (Figure 3D). This included, for example, the penicillin AMP (from an MVE of 42% to 25%), the cephalosporin ceftriaxone (CRO; 31% to 5%) the quinolone CIP (41% to 13%), and the aminoglycoside GEN (24% to 12%).

When eight AST results were included in the input – which corresponds to half of the complete diagnostic information in this study – thirteen of the sixteen antibiotics had an ME rate of less than 10%, and for ten antibiotics it was even lower than 5%. The VME rate for nine of the sixteen antibiotics was less than 10% and lower than 5% for four antibiotics. This was also reflected in the F1 scores which increased from 68% to 77% on average when going from four to eight input AST results. The overall high performance of the model in terms of F1 score, ME, and VME rates, suggested that it can be used to accurately predict complete susceptibility patterns for bacterial isolates.

Control of the major and very major error rates

In clinical practice, antibiotic treatment is typically based on the diagnostic tests of a single bacterial isolate. For predictions of diagnostic test results, the certainty of the predictions should be addressed as it could vary substantially. Indeed, if the uncertainty is too high, the prediction may need to be considered with care or completely disregarded. Therefore, we implemented a CICP algorithm that provides each prediction with a quantitative measure of uncertainty *(32)*. The certainty of each prediction was derived from a conformity measure, which in our case, was defined as the softmax score from the output of each antibiotic-specific neural network. The empirical distributions of softmax scores for both susceptible and resistant bacterial isolates were calculated for each of the antibiotics using a dedicated calibration dataset. These distributions were used to decide whether a new prediction was conformal to the susceptible and/or resistant isolates in the calibration dataset and, based on a pre-defined confidence level, sufficiently certain. The output of a prediction for one antibiotic is a prediction set, with either a single label, i.e. either susceptible or resistant, when there is enough conformity to only one of the softmax distributions; multiple labels, i.e. both susceptible and resistant, if there is enough conformity for both softmax distributions; or no label if there is no conformity to either group. The proportion of bacterial isolates that will, on average, be conformal to the correct label is governed by the confidence level $1 - \varepsilon$. Note that in this setting, ε corresponds to the average ME and the VME rates for susceptible and resistant bacterial isolates, respectively (see Methods for full details).

The empirically derived ME and VME rates were close to the pre-specified values of ε for all antibiotics (Figure 4A to D). For example, at a confidence level of 0.90 ($\varepsilon = 0.1$), the observed ME and VME rates were, on average, 9.8% (standard deviation 0.2%) and 10.4% (standard deviation 1.5%), respectively. The concordance between pre-specified and observed

error rates was sustained at higher confidence levels where, for 95% confidence, the average ME and VME rates were 4.9% (standard deviation 0.2%) and 5.3% (standard deviation 0.9%), respectively, and for 97.5% confidence, the average ME and VME rates were 2.47% (standard deviation 0.19%) and 2.54% (standard deviation 0.5%), respectively. The F1 score improved, as expected, with increasing confidence level, from an average value of 74% at 90% confidence to 85% and 92% for 95% and 97.5% confidence, respectively (Figure 4E). Thus, the results showed that the specified confidence levels calculated from empirical distributions were sufficiently stable between datasets.

The model could confidently predict the phenotype for a large majority of bacterial isolates and antibiotics

At a confidence level of 90%, 78% of the predictions were unambiguous and correct (only the correct label was predicted), but this number varied between antibiotics (Figure 5A). Cephalosporins, the quinolones LVX and OFX, and the aminoglycoside TOB all had a high proportion of correct unambiguous predictions (89.8% to 90.5%). However, the performance was lower for penicillins, the quinolones CIP and NAL, and the aminoglycoside GEN, reflecting a higher degree of uncertainty. The proportion of unambiguous predictions was, as expected, reduced when the confidence level was increased, from, 78.4% for 90% to 70.3% for 95% and, finally, to 59.8%, for 97.5%. On the other hand, the number of unambiguous predictions increased as more diagnostics information was provided to the model. At a 90% confidence level, the proportion of correct predictions with one single label increased from, on average, 72% when four AST results were included in the input, to 76%, 78%, 82%, and 84% when five, six, seven, and eight AST results were used, respectively (Figure 5B). The increase was especially large for the prediction of susceptibility to the penicillins AMP and AMX and for the prediction

of both susceptibility and resistance to the quinolone CIP. With eight input AST results, the model was able to make correct and unambiguous predictions for most of the remaining unknown antibiotics (84%) while maintaining both the major and very major error rates at 10%. This was also true, but to a lesser extent, for higher confidence levels where the proportion of unambiguous and correct predictions with eight input AST results were 78% and 68% for 95% and 97.5%, respectively.

DISCUSSION

In this study, we present a method that uses a transformer model to predict unavailable diagnostic information. When combined with conditional inductive conformal prediction to estimate the predictive uncertainty at the patient level, the method can abstain from making decisions unless the confidence is deemed sufficiently high. The model was applied to the diagnostics of infectious bacteria, a field that sees rapidly growing societal and economic costs due to the increasing challenges related to antibiotic resistance *(35, 36)*. We showed, in particular, that only a few antibiotic susceptibility test (AST) results are sufficient to accurately derive a more complete resistance profile. This has the potential to guide empirical antibiotic therapy for serious infections where the treatment needs to start before all diagnostic tests have finished.

The training of the method was done using a large and heterogeneous dataset, consisting of almost half a million bacterial isolates and more than three million AST results collected from thirty European countries. This dataset was compiled for surveillance purposes and contains AST results originally produced in routine diagnostics. Validation, where AST results were randomly excluded, showed that the model had generally high accuracy for the prediction of antibiotic

susceptibility. The performance improved further as results from more tests were included in the input data, demonstrating that the model could efficiently incorporate diagnostic information as it becomes available and, thus, over time, produce more certain predictions. Indeed, when eight of the sixteen antibiotics were used in the input, the model could predict most susceptibilities with a VME rate (false negative rate) as low as 5%. This shows that AI prediction can constitute a viable complement to laboratory diagnostics tests and could, potentially, be used to save time, reduce suffering, and lower economic costs.

Providing information about predictive uncertainty is essential in diagnostics where the knowledge at the population level is used to make predictions about individual patients. We addressed this challenge by implementing an algorithm based on conditional inductive conformal prediction *(32, 37, 38)* to provide each prediction with an accompanied confidence score. In our application, the pre-defined confidence corresponded to ME and VME rates, providing predictions that restrict the false positive and false negative rates to the desired levels. During the model testing, we found consistent error rates between datasets, and, for cephalosporins and quinolones, the confidence sets mainly contained a single label. However, a higher variation was observed for penicillins resulting in a larger proportion of ambiguous predictions (multiple labels). Furthermore, our implementation separates the uncertainty for susceptible and resistant predictions. This is valuable in the diagnostics of bacterial infections where a VME, i.e. the incorrect identification of a resistant isolate, may lead to ineffective antibiotic treatment. Therefore, VMEs are often considered to be the most serious errors, especially for life-threatening infections. The ability to set the confidence scores for susceptible and resistant predictions individually makes it, thus, possible to adjust the model to different clinical scenarios.

The model showed a clear difference in predictive performance between antibiotics, notoriously lower for penicillins, especially compared to cephalosporins. Historically, penicillins were one of the earliest classes of antibiotics and the first type of beta-lactam antibiotics to be introduced. Resistance against beta-lactam antibiotics is typically caused by enzymes that can break down the antibiotic using hydrolysis *(39)*. Penicillins have been widely used for over 80 years and there is, consequentially, a wide diversity of resistance genes that can be acquired by bacteria *(40)*. In contrast, resistance to cephalosporins – a later generation of beta-lactam antibiotics – is, to a larger extent, dependent on additional genetic events, such as mutations in chromosomal genes or the acquisition of broader spectrum resistance mechanisms *(39)*. These patterns were reflected in the data where resistance to penicillins was common. Indeed, as many as 89.8% of the bacterial isolates that were resistant to a single antibiotic were resistant to a penicillin antibiotic. Furthermore, it is plausible that the order of the evolution of multidrug-resistant bacteria has an impact on the performance of our model. Resistance phenotypes that are acquired initially will be harder to predict than those that commonly appear in later stages, simply due to the lack of other correlating susceptibilities. Results from additional diagnostic assays, such as targeted molecular tests detecting the presence of antibiotic resistance genes, could, thus, be a way to improve the performance further. Indeed, the flexibility of the transformers makes it possible to incorporate other types of diagnostic information, including genotypic information, as new words in the input sentence.

Finally, the AI methodology presented here has the potential to improve the diagnostics of infections caused by antibiotic-resistant bacteria. We argue that data-driven methods have the potential to replace selected diagnostics assays and thereby provide physicians with more comprehensive decision support at an earlier stage. This has the potential to improve the

treatment of bacterial infections and thereby decrease patient morbidity and mortality, reduce costs, and limit the overprescription of antibiotics.

MATERIALS AND METHODS

Data description

This study is based on data from The European Surveillance System (TESSy), which was collected as a part of the surveillance done by the European Centre for Disease Prevention and Control. The total dataset contains more than 9 million antibiotic susceptibility testing (AST) results for bacteria isolated from blood and cerebrospinal fluid from hospitalized patients in 30 different European countries. For each bacterial isolate, we retrieved the AST results together with the gender, sex, and age of the patient from which the bacterium was isolated, as well as the date of isolation and country where the test was performed. This metadata will be referred to collectively as the patient information. The analysis was limited to the susceptibility of *Escherichia coli* isolates, which was the most common species, from blood infections collected between 2013 to 2017. Tests that resulted in either resistant I or susceptible (S) were included, while tests with an intermediate (I) result were excluded. Only antibiotics with at least an 8% resistance rate were considered. This resulted in data covering sixteen antibiotics from four clinically relevant classes: aminoglycosides, cephalosporins, penicillins, and quinolones (Table 1). Furthermore, isolates with less than five tested antibiotics were removed. Also, if an antibiotic was tested multiple times for the same bacterial isolate, only the most recent test was included. The final dataset contained $n = 468{,}429$ bacterial isolates with 3,324,638 AST results, resulting in an average of 7.5 tests (standard deviation of 1.96) per isolate. After randomization,

the bacterial isolates were divided into three datasets: training data (80% of the isolates), calibration data (10%), and testing data (10%).

To increase the number of combinations of AST results, each dataset was expanded. For each bacterial isolate j, multiple data points were generated by randomly splitting the susceptibility test results into two groups. The first group together with the patent information, x_j, were considered known and, thus, used as input to the model. The second group, y_j, was hidden from the model and used to evaluate the predictive performance. The full details of the data expansion are provided in Supplemental Materials – Table S1. A made-up example of the information available for a bacterial isolate is presented below:

Available information for bacterial isolate j: "SV 30 M 2013_01 LVX_R AMC_S AMP_S TZP_R CTX_S GEN_S", represents a bacterium isolated at a Swedish hospital (SV) from a 30-year-old (30) male (M) patient in January 2013 (2013_01) where the isolated bacterium was tested against six antibiotics. The AST results indicated resistance to the antibiotics levofloxacin (LVX) and piperacillin/tazobactam (TZP), and susceptibility to amoxicillin/clavulanic acid (AMC), ampicillin (AMP), cefotaxime (CTX), and gentamicin (GEN). Two example data points, (x_k, y_k) and $(x_{k'}, y_{k'})$, that could, potentially, be created from this isolate:

$$(x_k, y_k) = (\text{“SV 30 M 2013_01 LVX_R AMC_S AMP_S TZP_R”}, \text{“CTX_S GEN_S”}),$$

$$(x_{k'}, y_{k'}) = (\text{“SV 30 M 2013_01 LVX_R AMC_S AMP_S CTX_S GEN_S”}, \text{“TZP_R”}).$$

Model description

Given a data point, (x_k, y_k), the model takes the input x_k and makes predictions $\hat{y}_k$ of the susceptibilities y_k. The input sentence is first complemented at the start with the classification word *CLS* and padded to a length $L = 19$ (the maximum length of a sentence) with *PAD* words if

needed. Each word is then converted into a linear embedding representation in the form of a d-dimensional vector that provides semantic meaning to the model ($d = 64$). These word embeddings are passed through $e = 1$ transformer encoder layers, each with one attention head followed by an add-and-normalize layer, a position-wise feed-forward layer (using 128 nodes), and, finally, another add-and-normalize layer. The first vector of the output from the encoder – representing the *CLS* word and containing information at the sentence level – is used as the input to $M = 16$ independent antibiotic-specific feed-forward networks, each of depth 2. The intermediate layer of the networks has 64 nodes and a rectified linear unit (ReLU) activation function followed by a normalization step, while the output vector of the final layers is a linear transformation to vectors of length 2 which was used to do binary classification. The isolate was classified as resistant or susceptible based on the largest output value.

The model was trained as follows. At each epoch, 300,000 bacterial isolates were randomly sampled from the training dataset and expanded as described above. The model was then trained on 512,000 randomly selected data points, divided into mini-batches of size 512. The loss was based on the cross entropy between the known (y_k) and predicted ($\hat{y}_k$) labels. The Adam optimizer was used to minimize the loss during 200 epochs using a fixed learning rate of 1×10^{-6}. The model was implemented and trained using Pytorch version 1.7.1.

Uncertainty control

An algorithm based on conditional inductive conformal prediction (CICP) was used to quantify the uncertainty of the predictions *(32)* with respect to the antibiotic and label (i.e. "susceptible" and "resistant"). Conformal prediction uses empirical data to estimate uncertainty and has previously been shown to suit complex diagnostic data for which valid distributional assumptions may be hard to make *(41)*. For a data point (x_k, y_k), the algorithm estimates a

prediction set Γ_k^{ε}, containing the predictions that are deemed sufficiently confident given a predefined confidence level $1 - \varepsilon$. The uncertainty for a prediction was based on its conformity measure, defined as the softmax transformation of the outputs of the neural networks. The conformity measure and, thus, the uncertainty, was derived individually for each antibiotic and each label (i.e. resistance or susceptible).

We estimated the empirical distributions for each conformity measure from the calibration dataset, which, for an antibiotic a, were assumed to contain $l = l_{a,S} + l_{a,R}$ data points, where $l_{a,S}$ and $l_{a,R}$ are the number of data points for susceptible and resistant bacterial isolates, respectively. For a data point (x_k, y_k), let $\alpha_k^{a,S}$ and $\alpha_k^{a,R}$ denote the softmax score for prediction of susceptibility and resistance to antibiotic a, respectively. The prediction sets were decided based on the empirical p-values $p_k^{a,S}$ and $p_k^{a,R}$, which were calculated according to

$$p_k^{a,S} = \frac{|i = 1, \ldots, l_{a,S} : \tilde{\alpha}_i^{a,S} \leq \alpha_k^{a,S}| + 1}{l_{a,S} + 1},$$

$$p_k^{a,R} = \frac{|i = 1, \ldots, l_{a,R} : \tilde{\alpha}_i^{a,R} \leq \alpha_k^{a,R}| + 1}{l_{a,R} + 1},$$

where $\tilde{\alpha}_i^{a,S}$ and $\tilde{\alpha}_i^{a,R}$ denotes the softmax scores calculated using the data points in the calibration dataset. At a confidence $1 - \varepsilon$, the prediction set was then formed by

$$\Gamma_k^{\varepsilon,a} = \{S \text{ if } p_k^{a,S} > \varepsilon\} \bigcup \{R \text{ if } p_k^{a,R} > \varepsilon\}.$$

Performance

The model's predictive performance was computed based on 853,466 training, 534,370 calibration, and 535,099 test data points. To evaluate the overall model performance, the F1 score was calculated. In addition, the *major error* (ME) rate, defined as the proportion of true

susceptible bacterial isolates being erroneously predicted, and the *very major error* (VME) rate, defined as the proportion of true resistant isolates being erroneously predicted, were also calculated. To measure the performance of the uncertainty control, true predictions encompassed prediction regions containing the true label for each antibiotic, and false predictions contained either no labels or only the wrong one. The evaluation of the prediction sets was based on 1,092,964 predictions. For comparison, a naive classifier was included. For antibiotic *a*, the naive classifier selected randomly between resistance and susceptibility with a probability equal to the proportion of bacterial isolates that were resistant to that antibiotic in the whole dataset.

ACKNOWLEDGMENTS

Data from The European Surveillance System – TESSy between 2013 to 2017, provided by Andorra, Albania, Armenia, Austria, Azerbaijan, Bosnia and Herzegovina, Belgium, Bulgaria, Belarus, Switzerland, Cyprus, Czechia, Germany, Denmark, Estonia, Greece, Spain, Finland, France, Georgia, Croatia, Hungary, Ireland, Israel, Iceland, Italy, Kyrgyzstan, Kazakhstan, Liechtenstein, Lithuania, Luxembourg, Latvia, Monaco, Republic of Moldova, Montenegro, Republic of North Macedonia, Malta, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Russian Federation, Sweden, Slovenia, Slovakia, San Marino, Tajikistan, Turkmenistan, Turkey, Ukraine, United Kingdom, Uzbekistan, Kosovo and released by The European Centre for Disease Prevention and Control (ECDC). The views and opinions of the authors expressed herein do not necessarily state or reflect those of ECDC. The accuracy of the authors' statistical analysis and the findings they report are not the responsibility of ECDC. ECDC is not responsible for conclusions or opinions drawn from the data provided. ECDC is not responsible for the correctness of the data and data management, data merging, and data collation

after the provision of the data. ECDC shall not be held liable for improper or incorrect use of the data.

DATA AVAILABILITY STATEMENT

The data used in this study belongs to the European Surveillance System – TESSy. We acknowledge the European Surveillance System – TESSy for data availability. The code used to generate the transformer model, train it and validate it, is available in the GitHub repository https://github.com/indajuan/Confidence-based-Prediction-of-Antibiotic-Resistance.

FUNDING

Swedish Research Council (VR), grant 2018-02835 (EK)

Swedish Research Council (VR), grant 2019-03482 (EK)

Chalmers AI Research Centre (CHAIR), (EK)

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTAL MATERIALS

Fig. S1. Performance of the model excluding patient information and of the naive classifier.

Table S1. Data expansion.

TABLES

Table 1. Bacterial isolates summary. The included antibiotics, the proportions of isolates tested and their susceptibility rate stratified on female and male patients, and the weights used in the loss function.

	Proportion of isolates tested[1]		**Loss function**
Antibiotic	**Female patients**	**Male patients**	**weights (S/R)[2]**
AMC – Amoxicillin/ clavulanic acid	45.4\% (72.2\% S)	47.0\% (68.9\% S)	0.45/0.55
AMP – Ampicillin	71.7\% (45.8\% S)	71.6\% (42.4\% S)	0.45/0.55
AMX – Amoxicillin	21.4\% (45.5\% S)	21.8\% (42.9\% S)	0.45/0.55
PIP – Piperacillin	9.6\% (52.5\% S)	9.1\% (49.2\% S)	0.45/0.55
TZP – Piperacillin/ tazobactam	49.0\% (93.3\% S)	50.0\% (91.9\% S)	0.15/0.85
CAZ – Ceftazidime	90.6\% (92.4\% S)	90.2\% (90.1\% S)	0.3/0.7
CRO – Ceftriaxone	29.1\% (91.3\% S)	27.4\% (88.9\% S)	0.3/0.7
CTX – Cefotaxime	80.3\% (90.0\% S)	81.0\% (87.1\% S)	0.3/0.7
FEP – Cefepime	29.1\% (92.5\% S)	30\% (90.1\% S)	0.3/0.7
CIP – Ciprofloxacin	93.9\% (82.5\% S)	94.5\% (75.8\% S)	0.3/0.7
LVX – Levofloxacin	24.8\% (80.0\% S)	23.6\% (72.0\% S)	0.3/0.7
MFX – Moxifloxacin	10.0\% (78.9\% S)	9.1\% (72.6\% S)	0.3/0.7
NAL – Nalidixic acid	7.2\% (76.0\% S)	8.3\% (71.6\% S)	0.45/0.55
OFX – Ofloxacin	8.1\% (83.4\% S)	9.1\% (78.6\% S)	0.3/0.7
GEN – Gentamicin	90.7\% (92.3\% S)	92.0\% (90.5\% S)	0.15/0.85
TOB – Tobramycin	46.4\% (91.6\% S)	47.4\% (89.4\% S)	0.15/0.85
Number of bacterial isolates	**249,025**	**219,404**	

[1]Proportion of bacterial isolates tested against each antibiotic, for bacteria isolated from female and male patients, respectively. The susceptibility rate for the isolates is shown in parentheses.

[2]Weight on the cross-entropy loss function for the two different labels: susceptible/resistant.

FIGURES

Fig. 1. The proposed model architecture and uncertainty control. (**A**) The architecture of the proposed model. The input sequence starts with a classification word, *CLS*, followed by the patient information (country, gender, age, date), and available antibiotic susceptibility data. The input is fixed to a length L through padding using the *PAD* word. We use a linear embedding to represent the input words numerically, which are fed into a transformer encoder. The *CLS* vector from the output of the encoder is then fed to *M* independent neural networks, each representing one antibiotic. The outputs of the neural networks are two-dimensional vectors, indicating susceptibility and resistance, respectively, that undergo a softmax rescaling. (**B**) Uncertainty control. A calibration dataset is used to build empirical distributions of conformity of resistant and susceptible predictions. The prediction regions for the test data are then created based on the deviation of the observed softmax score from the empirical distributions. See Methods for full details.

Fig. 2. Description of the dataset used for training. (**A**) The number of susceptible (opaque) and resistant (transparent) bacterial isolates tested per antibiotic. (**B**) The distribution of the number of antibiotics tested per bacterial isolate. (**C**) The proportion of bacterial isolates tested against each antibiotic for female (opaque) and male (transparent) patients. (**D**) The proportion of bacterial isolates susceptible to each antibiotic for female (opaque) and male (transparent) patients. The center line, box limits, and whiskers in C and D represent the median, upper, and lower quartiles, and 1.5 times the interquartile range, respectively.

Fig. 3. Performance of the transformer model. (**A**) Histogram of the number of antibiotic susceptibility testing (AST) results used as input for the model during training at each epoch. (**B-C**) Results from the training dataset (transparent) and testing dataset (opaque): F1 score, major

error (ME) rate, and very major error (VME) rate for the transformer model. (**D**) The predictive performance of the model as a function of the number of AST results included in the input (from transparent to opaque: 4 to 8 AST results).

Fig. 4. Empirical and expected error rates and performance of the model after CICP. For each antibiotic belonging to (**A**) penicillins, (**B**) cephalosporins, (**C**) quinolones, and (**D**) aminoglycosides. (**E**) Performance (F1 score, ME, and VME rates) as functions of the number of input AST results (from light to dark: 4 to 8 AST results) at a confidence level 0.9 ($\varepsilon = 0.1$).

Fig. 5. The proportion of correct predictions with respect to the number of labels in the prediction set. The proportion of correct predictions with a single label (opaque) and multiple labels (transparent), major errors (MEs), and very major errors (VMEs) with a single label (opaque) and empty set (transparent) predictions for resistant (R) and susceptible (S) isolates. (**A**) The proportions are shown for each antibiotic using three different confidence levels: 90%, 95%, and 97.5%. (**B**) The proportions are shown as a function of the number of input AST results (90% confidence level).

REFERENCES

1. Organization WH, others. 2015. Global action plan on antimicrobial resistance.

2. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, Han C, Bisignano C, Rao P, Wool E, others. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet 399:629–655.

3. O'Neill J. 2016. Tackling drug-resistant infections globally: final report and recommendations. Government of the United Kingdom.

4. Mercer DK, Torres MDT, Duay SS, Lovie E, Simpson L, von Köckritz-Blickwede M, de la Fuente-Nunez C, O'Neil DA, Angeles-Boza AM. 2020. Antimicrobial Susceptibility Testing of Antimicrobial Peptides to Better Predict Efficacy. Frontiers in Cellular and Infection Microbiology 10:326.

5. Jorgensen JH, Ferraro MJ. 1998. Antimicrobial susceptibility testing: general principles and contemporary practices. Clin Infect Dis 26:973–980.

6. Friedman ND, Temkin E, Carmeli Y. 2016. The negative impact of antibiotic resistance. Clin Microbiol Infect 22:416–422.

7. Bassetti M, Rello J, Blasi F, Goossens H, Sotgiu G, Tavoschi L, Zasowski EJ, Arber MR, McCool R, Patterson JV, others. 2020. Systematic review of the impact of appropriate versus inappropriate initial antibiotic therapy on outcomes of patients with severe bacterial infections. International journal of antimicrobial agents 56:106184.

8. Battle SE, Bookstaver PB, Justo JA, Kohn J, Albrecht H, Al-Hasan MN. 2016. Association between inappropriate empirical antimicrobial therapy and hospital length of stay in Gram-negative bloodstream infections: stratification by prognosis. Journal of Antimicrobial Chemotherapy 72:299–304.

9. Kumar A, Ellis P, Arabi Y, Roberts D, Light B, Parrillo JE, Dodek P, Wood G, Kumar A, Simon D, others. 2009. Initiation of inappropriate antimicrobial therapy results in a fivefold reduction of survival in human septic shock. Chest 136:1237–1248.

10. van den Bosch C, Hulscher ME, Akkermans RP, Wille J, Geerlings SE, Prins JM. 2017. Appropriate antibiotic use reduces length of hospital stay. Journal of Antimicrobial Chemotherapy 72:923–932.

11. Dolejska M, Papagiannitsis CC. 2018. Plasmid-mediated resistance is going wild. Plasmid 99:99–111.

12. Johnning A, Karami N, Tång Hallbäck E, Müller V, Nyberg L, Buongermino Pereira M, Stewart C, Ambjörnsson T, Westerlund F, Adlerberth I, Kristiansson E. 2018. The resistomes of six carbapenem-resistant pathogens - a critical genotype-phenotype analysis. Microb Genom 4.

13. Dias SP, Brouwer MC, van de Beek D. 2022. Sex and gender differences in bacterial infections. Infection and Immunity 90:e00283-22.

14. Murray KA, Preston N, Allen T, Zambrana-Torrelio C, Hosseini PR, Daszak P. 2015. Global biogeography of human infectious diseases. Proceedings of the National Academy of Sciences 112:12746–12751.

15. Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, Chodick G, Koren G, Shalev V, Kishony R. 2019. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nature Medicine 25:1143–1152.

16. Kanjilal S, Oberst M, Boominathan S, Zhou H, Hooper DC, Sontag D. 2020. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. Science Translational Medicine 12:eaay5067.

17. Stracy M, Snitser O, Yelin I, Amer Y, Parizade M, Katz R, Rimler G, Wolf T, Herzel E, Koren G, others. 2022. Minimizing treatment-induced emergence of antibiotic resistance in bacterial infections. Science 375:889–894.

18. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. 2021. The role of artificial intelligence in healthcare: a structured literature review. BMC Medical Informatics and Decision Making 21:1–23.

19. Yu K-H, Beam AL, Kohane IS. 2018. Artificial intelligence in healthcare. Nature Biomedical Engineering 2:719–731.

20. Benjamens S, Dhunnoo P, Meskó B. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ digital medicine 3:1–8.

21. Rajpurkar P, Chen E, Banerjee O, Topol EJ. 2022. AI in health and medicine. Nature Medicine 28:31–38.

22. The Medical Futurist. 2022. FDA-approved A.I.-based algorithms.

23. Begoli E, Bhattacharya T, Kusnezov D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence 1:20–23.

24. Kompa B, Snoek J, Beam AL. 2021. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine 4:1–6.

25. Feretzakis G, Loupelis E, Sakagianni A, Kalles D, Martsoukou M, Lada M, Skarmoutsou N, Christopoulos C, Valakis K, Velentza A, others. 2020. Using machine learning

techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in Greece. Antibiotics 9:50.

26. Feretzakis G, Sakagianni A, Loupelis E, Kalles D, Skarmoutsou N, Martsoukou M, Christopoulos C, Lada M, Petropoulou S, Velentza A, others. 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 27:214–221.

27. Devlin J, Chang M-W, Lee K, Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805.

28. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, others. 2020. Language models are few-shot learners. Advances in neural information processing systems 33:1877–1901.

29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser \Lukasz, Polosukhin I. 2017. Attention is all you need. Advances in neural information processing systems 30.

30. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589.

31. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nature biotechnology 1–3.

32. Vovk V. 2012. Conditional validity of inductive conformal predictors, p. 475–490. *In* Asian conference on machine learning. Proceedings of Machine Learning Research.

33. Vazquez J, Facelli JC. 2022. Conformal Prediction in Clinical Medical Sciences. Journal of Healthcare Informatics Research 1–12.

34. Reller LB, Weinstein M, Jorgensen JH, Ferraro MJ. 2009. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. Clinical infectious diseases 49:1749–1755.

35. European Centre for Disease Prevention and Control, European Medicines Agency. 2009. The bacterial challenge : time to react : a call to narrow the gap between multidrug-resistant bacteria in the EU and the development of new antibacterial agents. Publications Office.

36. Ahmad M, Khan AU. 2019. Global economic impact of antibiotic resistance: A review. Journal of Global Antimicrobial Resistance 19:313–316.

37. Vovk V, Gammerman A, Shafer G. 2005. Algorithmic Learning in a Random World. Springer-Verlag, Berlin, Heidelberg.

38. Papadopoulos H. 2008. Inductive Conformal Prediction: Theory and Application to Neural NetworksTools in Artificial Intelligence. IntechOpen, Rijeka.

39. Bush K, Bradford PA. 2020. Epidemiology of β-lactamase-producing pathogens. Clinical microbiology reviews 33:e00047-19.

40. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, Nykäsenoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L, Mossong J, Losch S, Ragimbeau C, Lund O, Aarestrup FM. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. Journal of Antimicrobial Chemotherapy 75:3491–3500.

41. Olsson H, Kartasalo K, Mulliqi N, Capuccini M, Ruusuvuori P, Samaratunga H, Delahunt B, Lindskog C, Janssen EA, Blilie A, others. 2022. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nature Communications 13:7761.

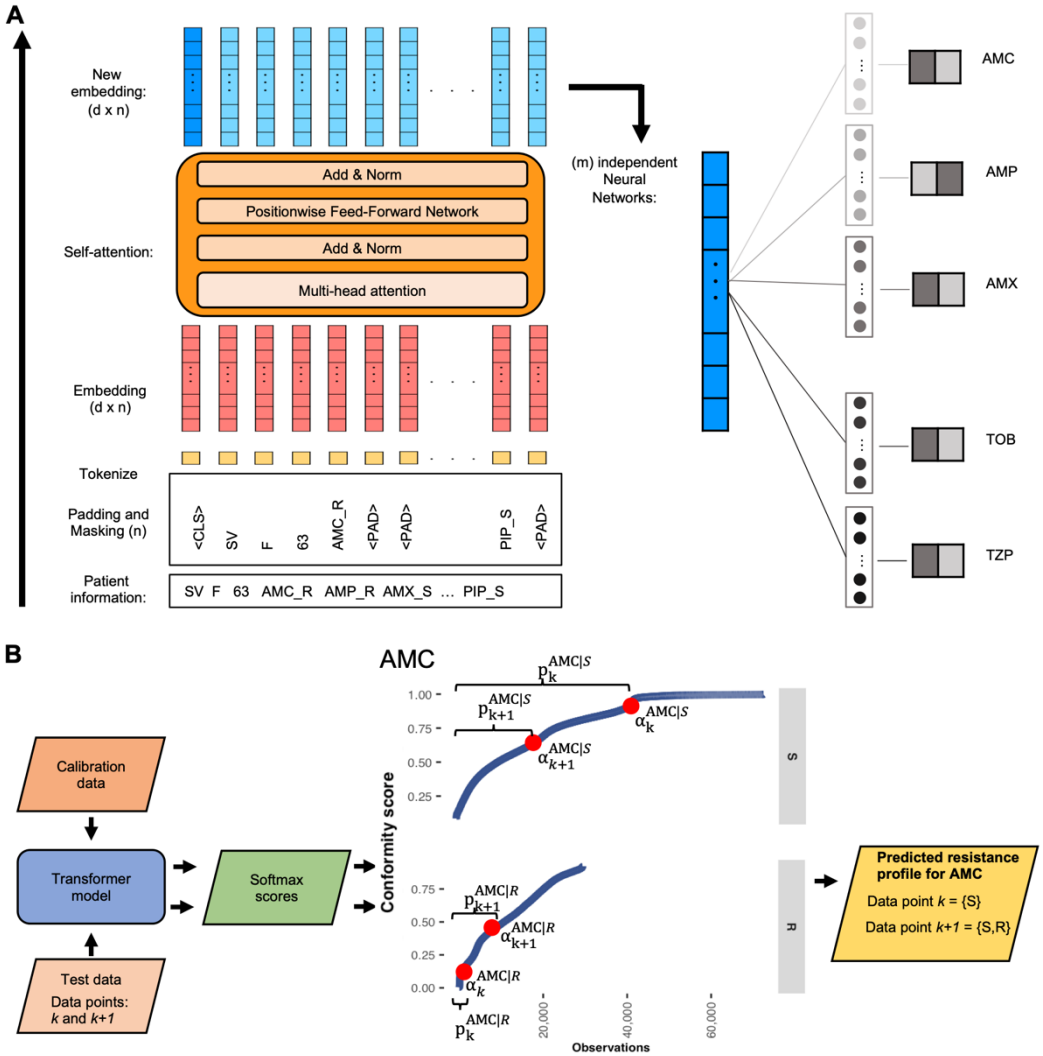

Fig. 1. The proposed model architecture and uncertainty control. (**A**) The architecture of the proposed model. The input sequence starts with a classification token, *CLS*, followed by the patient information (country, gender, age, date) and then available antibiotic susceptibility data. The input is fixed to a length L through padding using the *PAD* word. We use linear embedding to represent the input words numerically, which are fed into a transformer encoder. The *CLS* vector from the output of the encoder is then fed to *M*-independent neural networks, each representing one antibiotic. The outputs of the neural networks are two-dimensional vectors, indicating susceptibility and resistance, respectively, that undergo a softmax rescaling. (**B**) Uncertainty control. A calibration dataset is used to build empirical distributions of non-conformity of resistant and susceptible predictions. The prediction regions for the test data are then created based on the deviations from the empirical distributions, i.e. have a confidence score above a predefined cut-off. See Methods for full details.

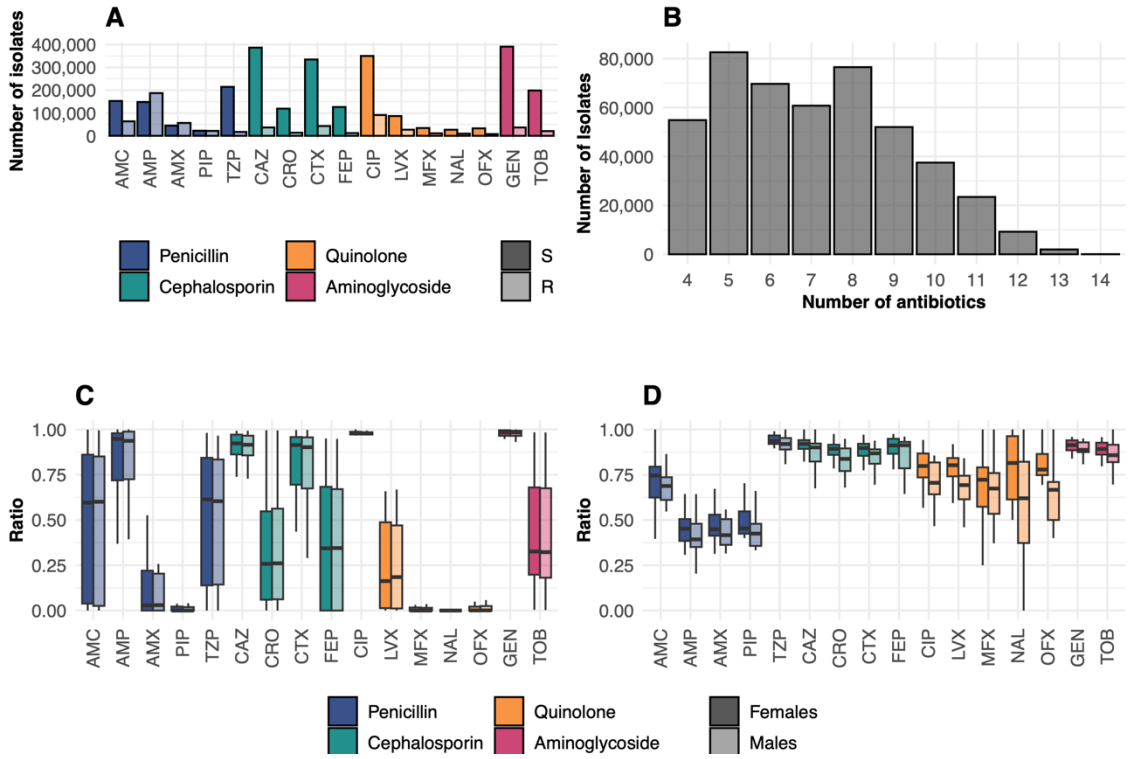

Fig. 2. Description of the dataset used for training. (**A**) The number of susceptible (opaque) and resistant (transparent) bacterial isolates tested per antibiotic. (**B**) The distribution of the number of antibiotics tested per bacterial isolate. (**C**) The proportion of bacterial isolates tested against each antibiotic for female (opaque) and male (transparent) patients. (**D**) The proportion of bacterial isolates susceptible to each antibiotic for female (opaque) and male (transparent) patients. The center line, box limits, and whiskers in C and D represent the median, upper and lower quartiles, and 1.5 times the interquartile range, respectively.

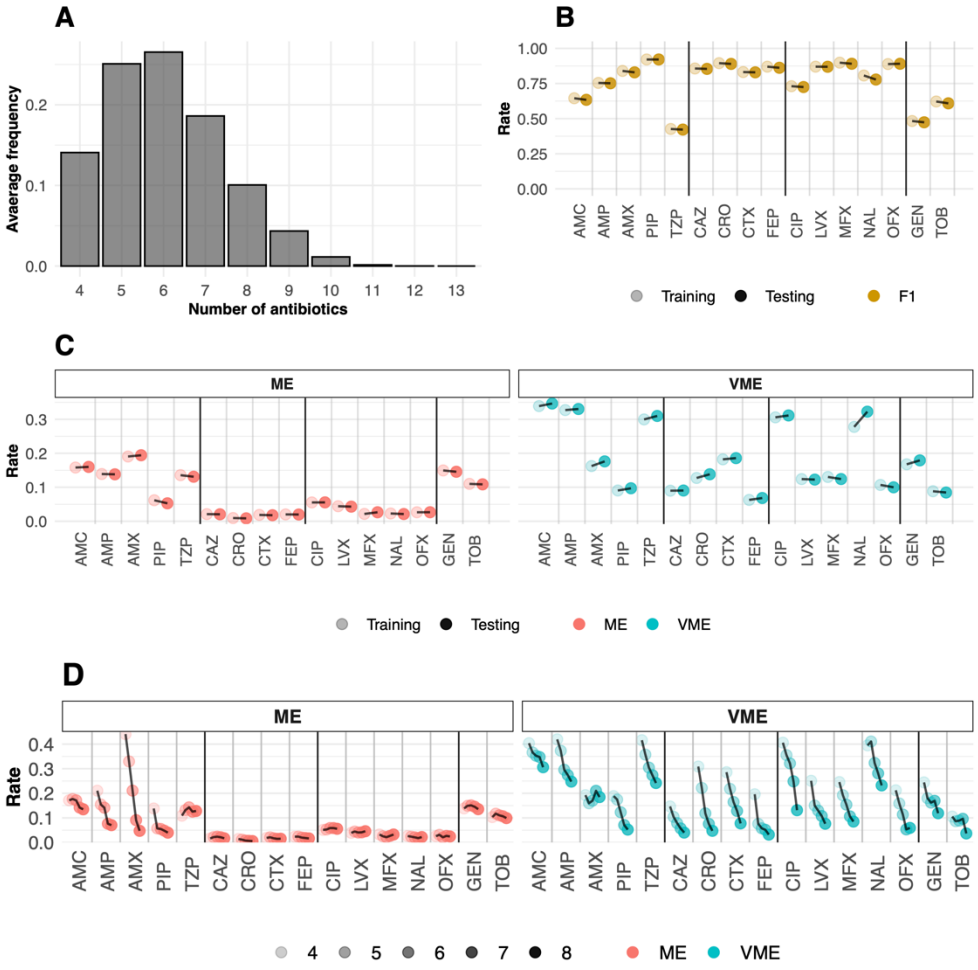

Fig. 3. Performance of the transformer model. (**A**) Histogram of the number of antibiotic susceptibility testing (AST) results used as input for the model during training at each epoch. (**B-C**) Results from the training dataset (transparent) and testing dataset (opaque): F1-score, major error (ME) rate, and very major error (VME) rate for the transformer model. (**D**) The predictive performance of the model as a function of the number of AST results included in the input (from transparent to opaque: 4 to 8 AST results).

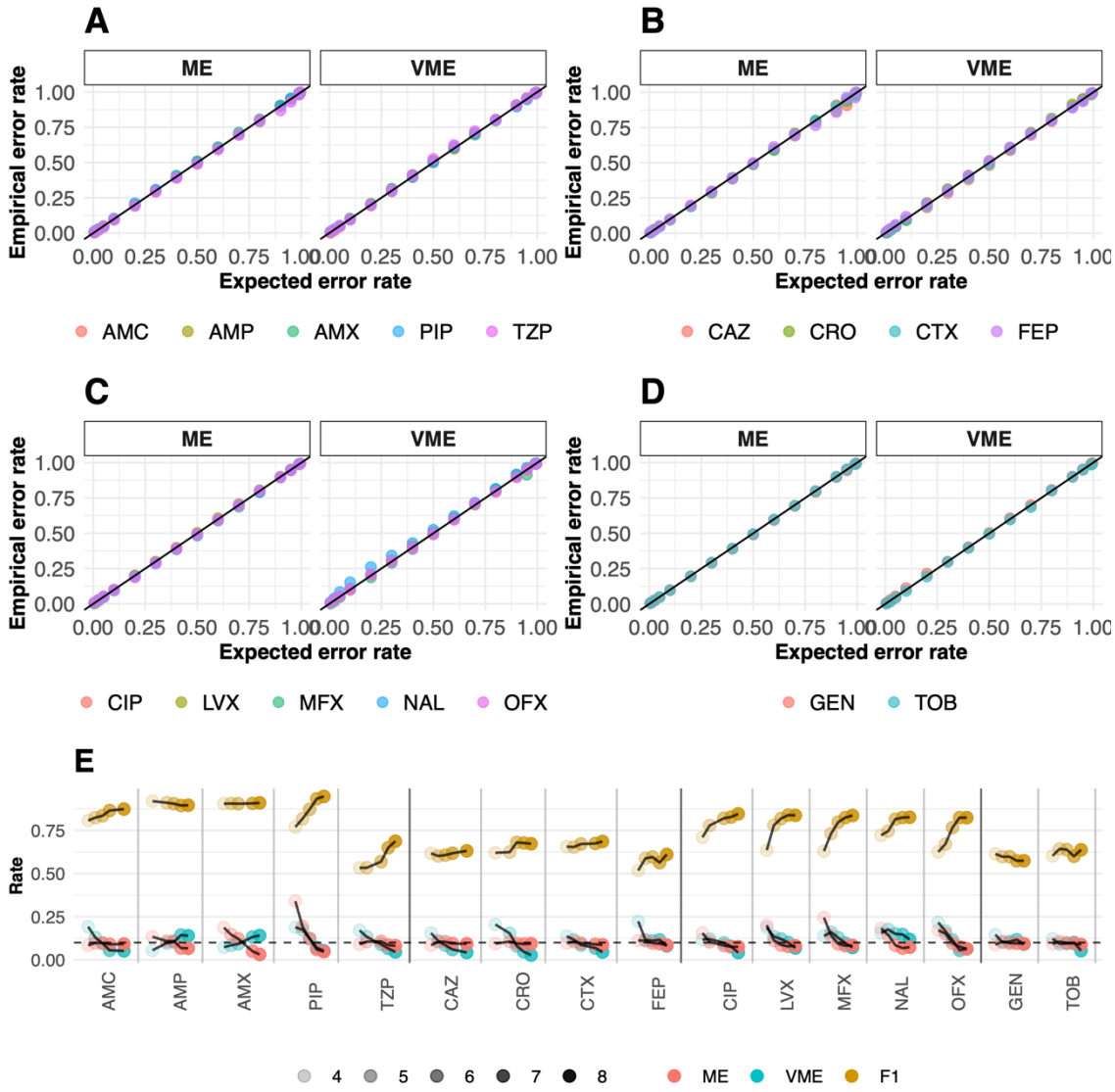

Fig. 4. Empirical and expected error rates and performance of the model after CICP. For each antibiotic by class for (**A**) penicillins, (**B**) cephalosporins, (**C**) quinolones, and (**D**) aminoglycosides. (**E**) Performance (F1-score, ME, and VME rates) as functions of the number of input AST results (from light to dark: 4 to 8 AST results) at a confidence level 0.9 ($\varepsilon = 0.1$).

Fig. 5. The proportion of correct predictions with respect to the number of labels in the prediction set. The proportion of correct predictions with a single label (opaque) and multiple labels (transparent), major errors (MEs), and very major errors (VMEs) with a single label (opaque) and empty set (transparent) predictions for resistant (R) and susceptible (S). (**A**) The proportions are shown for each antibiotic using three different confidence levels: 90%, 95%, and 97.5%. (**B**) The proportions are shown as a function of the number of input AST results (90% confidence level).