Confidence-based Prediction of Antibiotic Resistance at the Patient Level

Juan S. Inda-Díaz[1,2,3,*], Anna Johnning[1,2,4], Magnus Hessel[2,5], Anders Sjöberg[4,6], Anna Lokrantz[4], Lisa Helldal[5], Mats Jirstrand[4,6], Lennart Svensson[6] 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, Sweden.

[3]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Australia.

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

[5]Department of Clinical Microbiology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden.

[6]Department of Electrical Engineering, Chalmers University of Technology; Gothenburg, Sweden.

*Address correspondence to Juan S. Inda-Díaz, inda@chalmers.se and 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 seldom yield results before treatment needs to start, increasing patient risk and contributing to antibiotic overprescription. Here, we present a deep-learning method that leverages patient data and available AST results to predict antibiotic susceptibilities that have not yet been measured. After training on three million AST results from 30 European countries, the method achieved an average accuracy of 93% across bacterial species and antibiotics. It predicted susceptibility with an average major error rate below 5% for quinolones, cephalosporins, and carbapenems, and below 8% and 14% for aminoglycosides and penicillins, respectively. Furthermore, the model predicted resistance with an average very major rate below 10% for cephalosporins, carbapenems, and aminoglycosides, but with higher very major error rates for penicillins and quinolones. We combined the method with conformal prediction and demonstrated accurate estimation of the predictive uncertainty at the patient level. Our results suggest that AI-based decision support may offer new means to meet the growing burden of antibiotic resistance.

IMPORTANCE

Improved diagnostic tools are vital for maintaining efficient treatment of antibiotic-resistant bacteria and for reducing antibiotic overconsumption. In our research, we introduce a new deep learning-based method capable of predicting untested antibiotic resistance phenotypes. The method uses transformers, a powerful AI technique that efficiently leverages both antibiotic susceptibility tests (AST) and patient data simultaneously. The model produces predictions that can be used as time- and cost-efficient alternatives to results from cultivation-based diagnostic

assays. Significantly, our study highlights the potential of AI technologies to address the increasing prevalence of antibiotic-resistant bacterial infections.

INTRODUCTION

The global rise of antibiotic-resistant bacterial infections threatens human health globally *(1)*. In 2019, almost five million yearly deaths were attributed to antibiotic-resistant bacteria *(2)*, a number that is expected to continue to grow in the coming decades *(3)*. Reduced antibiotic efficacy in 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 effective 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 bacterial growth rate and the large number of antibiotics that may need to be tested for highly multidrug-resistant isolates. Tier-based testing strategies and time constraints usually yield incomplete AST results initially. For life-threatening infections, treatment needs to start as early as possible, often before AST result are available *(6)*. Under these circumstances, the choice of treatment is reduced to educated guesses based on limited diagnostic information *(7)*. This form of "empirical" treatment is associated with increased patient risks and overprescription of antibiotics (8–10).

65 Antibiotic resistance is commonly caused by resistance genes encoding various defense
66 mechanisms. These genes are often co-localized on mobile genetic elements, such as plasmids
67 and transposons *(11, 12)*. Multiple resistance genes can thus be transferred simultaneously
68 between bacterial cells, giving rise to strong correlations in susceptibility to different antibiotics.
69 The infecting bacterium and, therefore, its susceptibility profile are, furthermore, dependent on
70 patient characteristics, including age, sex, and the geographical region where the infection was
71 acquired (13–15). Indeed, patient data has previously been shown to contain valuable
72 information for selecting suitable antibiotic therapy for bacterial urinary tract infections (15–17).
73 There are, however, no AI-based methods that can combine patient data with the initial, often
74 incomplete AST results and thus guide treatment choice based on all available diagnostic
75 information. Indeed, integrating patient data with available AST results could enable more
76 accurate prediction of susceptibility to antibiotics that have not been tested, thereby providing
77 physicians with comprehensive diagnostic information at an earlier point in time.

78 Artificial intelligence (AI) and deep learning have been successfully applied to
79 diagnostics *(18, 19)*, but the focus has primarily been on image-based data commonly used in
80 radiology and pathology *(20)*. In contrast, methods for non-image multimodal data, which are
81 more prevalent in the diagnosis of infectious diseases, have received less attention *(21)*. Several
82 methods for predicting phenotypes from genotypic data have also been proposed (22), as have
83 machine-learning methods using electronic health records and patient data (23). There are,
84 however, few AI-based decision support systems for antibiotic treatment selection in use in
85 clinical settings *(24)*. A major culprit in the development of such methods is the complexity of
86 the diagnostic data, which is typically categorical (stratified test results and patient data) and
87 contains dependencies and redundancies. Also, when applied to antibiotic-resistant bacteria, the

88 incompleteness of the diagnostic data, particularly AST results, requires approaches that can
89 efficiently handle missing observations. Furthermore, since model accuracy depends on the
90 amount of available information, any prediction must be accompanied by estimates of its
91 uncertainty. Indeed, the possibility of disregarding predictions that are insufficiently confident is
92 vital for critical decision-making and, thus, essential for the adoption of AI-based methodologies
93 in healthcare settings *(25)*. However, today, most AI-based diagnostic methods are primarily
94 evaluated in populations and do not provide uncertainty information for patient-level predictions
95 *(26)*. Existing methods have, so far, been unable to adequately address these challenges and are
96 either limited to a few specific antibiotics, cannot incorporate missing AST data, or do not
97 provide any estimates of the predictive uncertainty *(27, 28)*.

98 In recent years, transformer-based models, such as BERT (bidirectional encoder
99 representations from transformers) *(29)* and GPT (generative pre-trained transformer) *(30)*, have
100 transformed natural language processing. These models operate on categorical input data, often
101 structured into sentences of words, and subject them to multi-head self-attention *(31)*. This
102 process enables inference of complex word dependencies directly from data and thereby predicts
103 missing parts. Therefore, we hypothesized that transformers may be suitable for predicting
104 antibiotic susceptibility results from a combination of incomplete diagnostic information and
105 patient data. Multimodal self-attention would enable the identification of complex dependencies
106 among diagnostic data types and facilitate extrapolation to susceptibilities that have not been
107 tested. Transformers have previously been shown to be highly useful beyond natural language
108 processing *(32, 33)*, but are rarely used for diagnosing infectious diseases.

109 In this study, we present a novel transformer-based method that accurately predicts
110 antibiotic susceptibilities using patient data and incomplete diagnostic information for three

common bacterial pathogens, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. We combine the model with conditional inductive conformal prediction *(34)* to estimate the uncertainty of each prediction at the patient level, thereby enabling dismissal of predictions with too low certainty. The model was trained and evaluated on a large, heterogeneous dataset containing AST results from blood infections collected during routine care across 30 European countries. Our results showed that the model could accurately predict susceptibility to a wide range of antibiotics, even 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 additional AST results become available. Finally, we demonstrate that predictions can be made within predefined confidence levels for most bacterial species and antibiotics, which allows control of both the major and very major error rates. We conclude that the combination of transformers and conditional inductive conformal predictions constitutes an appealing class of models for integrating and predicting 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 set 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 has available results. The model uses two transformers, which summarize the patient and AST information in two sentence-specific classification (CLS) vectors, one representing patient-specific information and one representing bacterial isolate-specific information. These vectors

134 then serve as input to multiple antibiotic-specific feed-forward neural networks, each predicting
135 the probability of susceptibility to the corresponding antibiotic. The uncertainty is estimated
136 using inductive conformal prediction, conditioned on the susceptibility; hence, controlling the
137 false-positive (major error) and false-negative (very major error) rates for each antibiotic *(34,*
138 *35)*. The full details of the model architecture and the uncertainty estimation are provided in
139 Methods.

140 The model was trained and evaluated on data from The European Surveillance System
141 (TESSy) (https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy),
142 which contains AST results for 1,161,303 *E. coli*, 301,506 *K. pneumoniae*, and 166,448 *P.*
143 *aeruginosa* isolates collected from blood infections of patients in 30 European countries. The
144 training and evaluation were performed on 20 (16, 18, and 11 for *E. coli*, *K. pneumoniae*, and *P.*
145 *aeruginosa,* respectively) commonly used antibiotics that belong to five large and clinically
146 relevant antibiotic classes: aminoglycosides, carbapenems, cephalosporins, penicillins, and
147 quinolones (Table 1). A bacterial isolate was, on average, tested for susceptibility to 7.3, 8.8, and
148 7.8 antibiotics, with standard deviations of 1.9, 2.4, and 1.7 for *E. coli*, *K. pneumoniae*, and *P.*
149 *aeruginosa*, respectively (Figures S1 and S2). The most commonly tested antibiotics were
150 ceftazidime (CAZ) and ciprofloxacin (CIP), for which at least 93% of isolates were tested,
151 regardless of pathogen, country of origin, or patient gender. In contrast, less than 10% of the *E.*
152 *coli* and *K. pneumoniae* isolates were tested for piperacillin (PIP), moxifloxacin (MFX), and
153 nalidixic acid (NAL, Figure S3). The susceptibility rate for *E. coli* isolates was lowest for the
154 penicillins amoxicillin (AMX), ampicillin (AMP), and PIP (43%–50%), whereas for *K.*
155 *pneumoniae* isolates it was lowest for PIP (20%). For other antibiotics and *P. aeruginosa*
156 isolates, the susceptibility rates were higher, representing a more unbalanced dataset. A slightly

157 higher susceptibility rate was observed among isolates collected from female patients (Figure
158 S4).

159 **Predictions of antibiotic susceptibility have high performance**

160 To assess how well the model predicts the susceptibility of antibiotics not included as
161 inputs, we used 70% of the bacterial isolates for 5-fold cross-validation and model training. We
162 used 15% of the isolates for testing, and the remaining 15% for calibration of conformal
163 prediction. During training, calibration, and testing, we selected a random number of antibiotics
164 as input to the model (distribution available in Figure S5; mean 5.7, 6.3, and 5.7 with standard
165 deviations of 1.3, 1.5, and 1.1 for *E. coli*, *K. pneumoniae*, and *P. aeruginosa* isolates,
166 respectively; see Methods for details) together with patient data. The susceptibility of the
167 remaining tested antibiotics was assumed to be unknown and, therefore, masked from the input
168 data. The model's predictions were then compared with the antibiotics masked from the input to
169 evaluate performance. In this setting, the model had an overall high performance that did not
170 differ substantially between folds in the cross-validation, with a standard error (se) below 0.015
171 for all antibiotics and bacterial species except the major error (ME) for the antibiotic PIP in *K.*
172 *pneumoniae* (se = 0.016 for both training and test datasets). The performance was largely
173 consistent across the training and test processes (Figures 2 and S6). There were, however,
174 apparent differences in performance between the three bacterial species and between antibiotics.
175 The F1 score (the harmonic mean of precision and recall) in the test dataset for *E. coli and K.*
176 *pneumoniae* isolates was highest for cephalosporins (91% and 96% on average, respectively),
177 and quinolones (86% and 93% on average, respectively), though lower for penicillins (74% and
178 88% on average, respectively, Figure 2A). For *P. aeruginosa* isolates, the F1 score was 84% for
179 penicillins, and 82% for quinolones and cephalosporines. Carbapenems had an average F1 score

180 of 90% and 76% for *K. pneumoniae* and *P. aeruginosa* isolates, respectively, while
181 aminoglycosides had the lowest F1 scores: 65% for *E. coli* and 76% for *K. pneumoniae* and *P.*
182 *aeruginosa* isolates.

183 Next, we evaluated the model based on the ME rate, defined as the proportion of
184 susceptible bacterial isolates erroneously predicted as resistant, and the very major error (VME)
185 rate, defined as the proportion of resistant isolates erroneously predicted as susceptible. The ME
186 and VME rates are standard performance measures in antibiotic susceptibility testing, and are
187 frequently used to evaluate and compare diagnostic methods *(36)*. Based on the test dataset,
188 cephalosporins had, on average over all predictions in the test data, an ME rate of 1.4% (1%–
189 1.8%), 2.3% (1.5%–3.9%), and 5.2% (5.1%–5.2%) for *E. coli*, *K. pneumoniae*, and *P.*
190 *aeruginosa* isolates, respectively, while the corresponding VME rate was, on average, 6.7%
191 (5.6%–9.4%), 2.3% (1.2%–3.2%), and 12.2% (9.7%–12.8%) (Figure 2B). For quinolones, the
192 average ME rate was 2.2% (2%–5%), 2.9% (1.6%–4.7%), and 3.3% (2.5%–4.2%), while the
193 VME rate was, on average 19.2% (8.9%–28.6%), 8.3% (4.2%–13.4%), and 21.9% (20.3%–
194 23.5%) for *E. coli*, *K. pneumoniae*, and *P. aeruginosa* isolates, respectively. For *E. coli* isolates,
195 the penicillins had a higher overall average ME rate of 12.5% (4.1%–16.6%). Here, PIP had the
196 lowest VME rate (12.2%), while the other penicillins had rates between 23% and 26%. For *K.*
197 *pneumoniae* and *P. aeruginosa* isolates, the average ME for penicillins, excluding PIP, was 7.3%
198 and 6.3%, respectively, while the VME was 11%. We noted that the model struggled to predict
199 PIP resistance in *K. pneumoniae* with an ME as high as 51%. This may be a consequence of data
200 imbalance, where *K. pneumoniae* had a resistance rate of 80% to PIP (compared to 50% and 22%
201 for *E. coli* and *P. aeruginosa*, respectively) but accounted for only 11% of the data for this
202 antibiotic. Aminoglycosides had average ME/VME rates of 7.3/16.3, 10.2/7.4, and 5.5/9.5, for *E.*

203 *coli, K. pneumoniae*, and *P. aeruginosa* isolates, respectively. Finally, the ME and VME rates for
204 carbapenems for *K. pneumoniae* isolates were 2.3% (1.7%–3.1%) and 4.4% (2.8%–7.3%),
205 respectively, and for *P. aeruginosa* isolates, 8.4% (5.3%–11.4%) and 14.6% (10.9%–18.4%),
206 respectively.

207 The number of available AST results influenced the model's performance. When the
208 number of AST results used as input to the model increased from four to eight antibiotics,
209 significant reductions were seen in the ME rate for penicillins, where AMX and AMP dropped
210 from 34% and 22% to 3% and 7%, respectively, and the VME of CIP, which fell from 36% to
211 7% in *E. coli* isolates (Figure 3D). Interestingly, the drop for other antibiotics and pathogens was
212 more modest or non-existent. A substantial reduction in the VME rate was also observed across
213 most antibiotics and pathogens.

214 When eight AST results were included in the input, fourteen of sixteen antibiotics in *E.*
215 *coli* isolates, sixteen of twenty antibiotics in *K. pneumoniae*, and all antibiotics in *P. aeruginosa*
216 had an ME rate below 10%, and below 5% for eleven, thirteen, and ten antibiotics in each
217 pathogens, respectively. Similarly, the VME rate was below 10% for nine, sixteen, and ten, and
218 below 5% for six, twelve, and three of the antibiotics corresponding to each pathogen. Going
219 from four to eight input AST results increased the F1 scores from 75% to 87% for *E. coli*, from
220 85% to 92% for *K. pneumoniae*, and from 75% to 84% for *P. aeruginosa*, on average. The
221 model's overall high performance, as indicated by its F1 score, ME, and VME rates, suggests that
222 it can be used to predict complete susceptibility patterns for bacterial isolates accurately.

223 **Control of the major and very major error rates at the patient level**

224 In clinical practice, decisions are taken at the patient level. It is thus vital that decision
225 support based on predictions also conveys information about the confidence in the prediction.

226 Indeed, if the uncertainty is too high, the prediction may need to be considered with care or
227 completely dismissed until more diagnostic information becomes available. Therefore, we
228 implemented an algorithm that assigns a quantitative measure of uncertainty to each individual
229 prediction *(34)*. The algorithm was based on conditional inductive conformal prediction, where
230 the certainty of each prediction was derived from a conformity measure, defined in our case as
231 the softmax score from each antibiotic-specific neural network (Figure 1B). The distributions of
232 softmax scores for both susceptible and resistant bacterial isolates were empirically calculated
233 for each antibiotic and each of the three pathogens, using a dedicated calibration dataset. These
234 distributions were used to determine whether a new prediction aligns with the susceptible and/or
235 resistant isolates in the calibration dataset and, based on a predefined confidence level, is deemed
236 sufficiently certain. The output of a final prediction for one antibiotic is a prediction set, with
237 either a) a single label, i.e., either susceptible or resistant, when there is enough conformity to
238 only one of the softmax distributions; b) multiple labels, i.e. both susceptible and resistant, if
239 there is enough conformity for both softmax distributions; or c) no label if there is no conformity
240 to either group. The proportion of predictions that will, on average, have the correct label in their
241 prediction set is governed by the confidence level $1 - \varepsilon$. Note that in this setting, ε corresponds
242 to the average ME and the VME rates for susceptible and resistant bacterial isolates for each
243 pathogen, respectively (see Methods for full details).

244 The empirically derived ME and VME rates were close to the prespecified values of ε for
245 all antibiotics (Figures S7-S11). For example, at a confidence level of 90% ($\varepsilon = 0.1$), the
246 observed ME rates were, on average, 10.1%, 9.9%, and 9.9% (standard deviation 0.4%, 1.3%,
247 and 0.3%), and the observed VME rates were, on average 9.7%, 9.4%, and 9.3% (standard
248 deviation 1.43%, 0.95%, and 1.7%) for *E. coli*, *K. pneumoniae*, and *P. aeruginosa* isolates,

249 respectively. The concordance between pre-specified and observed error rates was sustained at
250 higher confidence levels where, for a confidence level of 95%, the average ME rates were 5.1%,
251 5.1%, and 5% (standard deviation 0.5%, 1%, and 0.3%) and the average VME rates were 5%,
252 4.7%, and 5% (standard deviation 0.8%, 0.6%, and 1.4%) for *E. coli*, *K. pneumoniae*, and *P.*
253 *aeruginosa* isolates, respectively. Thus, the results showed that the specified confidence levels
254 calculated from empirical distributions were sufficiently stable between datasets.

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

257 At a confidence level of 90%, as many as 82%, 89%, and 86% of the predictions were
258 unambiguous and correct (only the correct label was in the prediction set) for *E. coli*, *K.*
259 *pneumoniae*, and *P. aeruginosa* isolates, respectively; however, this number varied between
260 antibiotics (Figure 4A, 5A, 6A). Cephalosporins had a high proportion of correct unambiguous
261 predictions (90%) across the three pathogens, and quinolones had 90% for *K. pneumoniae*. For
262 the remaining antibiotics, between 82% and 89% of the predictions were correct and
263 unambiguous, except for penicillins for *E. coli* (71%) and quinolones for *P. aeruginosa*, which
264 showed higher uncertainty. The proportion of unambiguous predictions was, as expected,
265 reduced when the confidence level was increased; to 78%, 89%, and 77% at a 95% confidence
266 level and, finally, to 72%, 85%, and 63% at a 97.5% confidence level for *E. coli*, *K. pneumoniae*,
267 and *P. aeruginosa*, respectively. On the other hand, the number of unambiguous predictions
268 increased as more diagnostic information was provided to the model. At a 90% confidence level,
269 the proportion of correct unambiguous predictions increased from, on average, 76%, 84%, and
270 83% when four AST results were included in the input, to 89%, 91%, and 91% when eight AST
271 results were included in the input for *E. coli, K. pneumoniae*, and *P. aeruginosa* isolates,

272 respectively (Figure 4B, 5B, 6B). The increase was substantial for predicting susceptibility to the
273 penicillins AMP and AMX in *E. coli* isolates, to PIP and OFX in *K. pneumoniae* isolates, and for
274 predicting susceptibility and resistance to CIP, as well as susceptibility to LVX and IPM in *P.*
275 *aeruginosa* isolates. With eight input AST results, the model achieved 90% correct,
276 unambiguous predictions for most of the remaining masked antibiotics. This was also true for
277 higher confidence levels, where the proportion of unambiguous and correct predictions with
278 eight input AST results was 89% and 84% for 95% and 97.5% confidence, respectively.

279

280 **DISCUSSION**

281 In this study, we present a method that uses a transformer model to predict unavailable
282 diagnostic information. When combined with conditional inductive conformal prediction to
283 estimate the predictive uncertainty at the patient level, the method can abstain from making
284 decisions unless the confidence is deemed sufficiently high. The model was applied to the
285 diagnostics of infectious bacteria, a field facing rapidly rising societal and economic costs due to
286 the growing challenges related to antibiotic resistance *(37, 38)*. We showed that only a few
287 antibiotic susceptibility test (AST) results are sufficient to accurately derive a more complete
288 resistance profile. This has the potential to guide antibiotic therapy for serious infections, where
289 treatment must begin before all diagnostic tests are completed.

290 The method was trained on a large, heterogeneous dataset comprising over 12 million
291 AST results from more than a million bacterial isolates of three pathogens collected from 30
292 European countries. This dataset was compiled for surveillance purposes and contains AST
293 results originally produced in routine diagnostics. Validation, in which AST results were
294 randomly excluded, showed that the model had generally high accuracy in predicting antibiotic

295 susceptibility. The performance improved further as more ASTs were included in the input data,
296 demonstrating that the model could efficiently incorporate diagnostic information as it becomes
297 available and, thus, over time, produce more certain predictions. Indeed, when AST for eight
298 antibiotics was used as input, the model could predict most susceptibilities with a VME rate
299 (false-negative rate) as low as 10%. This indicates that AI prediction can serve as a viable
300 complement to laboratory diagnostic tests and may be utilized to save time, alleviate suffering,
301 and reduce economic costs.

302 Providing information about predictive uncertainty is crucial in diagnostics, where
303 population-level knowledge is used to predict outcomes for individual patients. We addressed
304 this challenge by implementing an algorithm based on conditional inductive conformal
305 prediction *(34, 39, 40)* to provide each prediction with an accompanying confidence score. In our
306 application, the predefined confidence corresponded to ME and VME rates, yielding predictions
307 that limit false-positive and false-negative rates to the desired levels. During model testing, we
308 found consistent error rates across datasets and, for cephalosporins and quinolones, the
309 prediction sets mainly contained a single, correct label. However, a higher variation was
310 observed for penicillins, resulting in a larger proportion of ambiguous predictions (multiple
311 labels). Furthermore, our implementation separates the uncertainty for susceptible and resistant
312 predictions. This is valuable in the diagnostics of bacterial infections, where a VME, i.e., the
313 incorrect identification of a resistant isolate, may lead to ineffective antibiotic treatment.
314 Therefore, VMEs are often considered to be the most serious errors, especially for life-
315 threatening infections. The ability to set confidence scores for susceptible and resistant
316 predictions individually enables the model to be adjusted for different clinical scenarios.

317 The model showed a clear difference in predictive performance between antibiotics,
318 notoriously lower for penicillins, especially compared to cephalosporins. Historically, penicillins
319 were among the earliest classes of antibiotics and the first beta-lactam antibiotics introduced.
320 Resistance to beta-lactam antibiotics is typically caused by enzymes that hydrolyze the
321 antibiotics *(41)*. Penicillins have been widely used for over 80 years, and bacteria can
322 consequently acquire a wide diversity of resistance genes *(42)*. In contrast, resistance to
323 cephalosporins – a later generation of beta-lactam antibiotics – is, to a larger extent, dependent
324 on additional genetic events, such as mutations in chromosomal genes or the acquisition of
325 broader-spectrum resistance mechanisms *(41)*. These patterns were reflected in the data, where
326 resistance to penicillins was common. Indeed, as many as 87.7% of the *E. coli* and 61% of the *K.*
327 *pneumoniae* bacterial isolates resistant to a single antibiotic were resistant to a penicillin.
328 Furthermore, it is plausible that the order of evolution of multidrug-resistant bacteria affects the
329 performance of our model. We also consider resistance phenotypes that are initially acquired to
330 be harder to predict than those that commonly appear in later stages of evolution, simply because
331 there are no other correlating susceptibilities. Results from additional diagnostic assays, such as
332 targeted molecular tests that detect antibiotic resistance genes and genomic data, could further
333 improve performance. Indeed, the flexibility of the transformers allows incorporating other types
334 of diagnostic information, including genotypic information, as new words in the input sentence.

335 Ultimately, the AI methodology presented here has the potential to enhance the diagnosis
336 of infections caused by antibiotic-resistant bacteria. We argue that data-driven methods have the
337 potential to replace selected diagnostic assays, thereby providing physicians with more
338 comprehensive decision support at an earlier stage. This has the potential to improve the

339 treatment of bacterial infections, thereby decreasing patient morbidity and mortality, reducing
340 costs, and limiting the overprescription of antibiotics.
341

342 **MATERIALS AND METHODS**

343 **Data description**

344 This study is based on data from the European Surveillance System (TESSy), collected as
345 part of surveillance conducted by the European Centre for Disease Prevention and Control. The
346 dataset contains more than 12 million antibiotic susceptibility testing (AST) results for bacteria
347 isolated from blood and cerebrospinal fluid of hospitalized patients across 30 European
348 countries. For each bacterial isolate, we retrieved the species name, AST results, the patient's
349 gender and age, the date of bacterial isolation, and the country where the test was performed.
350 This metadata will be referred to collectively as the patient data. The analysis was limited to the
351 susceptibility of *Escherichia coli, Klebsiella pneumoniae,* and *Pseudomonas aeruginosa* isolates,
352 which were the most common species, from blood infections collected between 2007 and 2020.
353 Tests that resulted in either a resistant (R) or a susceptible (S) result were included, while tests
354 with an I result were excluded. Only antibiotics with a resistance rate of at least 8% were
355 considered. This resulted in data covering sixteen antibiotics from five clinically relevant classes:
356 aminoglycosides, cephalosporins, penicillins, quinolones, and carbapenems (Table 1).
357 Furthermore, isolates tested for fewer than five antibiotics were removed. If an antibiotic was
358 tested multiple times for the same bacterial isolate, only the most recent test was included. The
359 final dataset contained $n = 1{,}629{,}257$ bacterial isolates with 12,433,856 AST results, resulting
360 in an average of 7.6 tests (standard deviation of 2.1) per isolate. After randomization, the
361 bacterial isolates were divided into three datasets: a training dataset (70%), a calibration dataset

362 (15%), and a test dataset (15%). The train dataset was further divided into five groups for 5-fold
363 cross-validation.

364 **Data expansion**

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

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

380 $$(x_k, y_k) = (\text{"ESCCOL SV 30 M 2013_01 LVX_R AMC_S AMP_S TZP_R"}, \text{"CTX_S GEN_S"}),$$

381 $$(x_{k'}, y_{k'}) = (\text{"ESCCOL SV 30 M 2013_01 LVX_R AMC_S AMP_S CTX_S GEN_S"}, \text{"TZP_R"}).$$

382 **Model Architecture and Training**

383 Given a data point, (x_k, y_k), the model takes the input x_k and makes predictions $\hat{y}_k$ of the
384 susceptibilities y_k. The input sentence x_k is split into patient data and AST data, both of which
385 contain the word for the bacterial species, and both are complemented at the start with the
386 classification word, *CLS.* The patient data and the AST data are padded to a length $L = 6$ and
387 $L = 21$ with *PAD* words if needed. Each word is then converted into a linear embedding
388 representation in the form of a d-dimensional vector that provides semantic meaning to the
389 model ($d = 128$). These word embeddings are passed through two transformer encoder layers,
390 each with two attention heads, followed by an add-and-normalize layer, a position-wise feed-
391 forward layer (using 256 nodes), and, finally, another add-and-normalize layer. The first vector
392 of the output from each encoder– representing the *CLS* word and containing information at the
393 sentence level– are concatenated and fed to a linear transformation. The output is used as input to
394 $M = 20$ independent antibiotic-specific feed-forward networks, each of depth two. The
395 intermediate layer of the network has 128 nodes with rectified linear unit (ReLU) activation
396 function, followed by a normalization step, while the final layer outputs a linear transformation
397 to vectors of length two, which are used for binary classification. The isolate was classified as
398 resistant or susceptible based on the largest output value. Additionally, models with the
399 combinations of 1 or 2 encoder heads and layers, with word embedding of size 64 or 128, and a
400 position-wise feed-forward layer of 128 or 256, with patient data and no patient data, were
401 trained (Figure S13-S15). The proposed model achieved higher accuracy, lower ME, and lower
402 VME in 23, 14, and 10 out of 45 pathogen-antibiotic combinations, respectively. The
403 combination of one head, one layer, with word embedding of size 128 and a position-wise feed-
404 forward layer of size 256 was the only one to have more pathogen-antibiotic combinations with
405 lower VME (13).

406 The model was trained as follows. At each epoch, 248,000 bacterial isolates were
407 randomly sampled from the train dataset and expanded as described in Methods – Data
408 expansion. The model was then trained on 512,000 randomly selected data points from the data
409 expansion step, divided into mini-batches of size 512. The loss was based on the cross-entropy
410 between the known (y_k) and predicted ($\hat{y}_k$) labels. The Adam optimizer was used to minimize
411 the loss over 700 epochs with a fixed learning rate of 1×10^{-6}. The model was implemented and
412 trained using Pytorch version 1.7.1.

413 **Uncertainty control**

414 An algorithm based on conditional inductive conformal prediction was used to quantify
415 the uncertainty of the predictions *(34)* with respect to the antibiotic and label (i.e., "susceptible"
416 and "resistant") (34). Conformal prediction uses empirical data to estimate uncertainty and has
417 previously been shown to suit complex diagnostic data for which valid distributional
418 assumptions may be hard to make *(43)*. For a data point (x_k, y_k) belonging to pathogen q, the
419 algorithm estimates a prediction set $\Gamma_k^{\varepsilon,q}$, containing the predictions that are deemed sufficiently
420 confident given a predefined confidence level $1 - \varepsilon$. The uncertainty for a prediction was based
421 on its conformity measure, defined as the softmax transformation of the outputs of a neural
422 network. The conformity measure and, thus, the uncertainty, was derived individually for each
423 antibiotic and each label (i.e., resistance or susceptible).

424 We estimated the empirical distributions for each conformity measure from the
425 calibration dataset separately for each pathogen, which, for an antibiotic a and pathogen q, were
426 assumed to contain $l_{q,a} = l_{q,a,S} + l_{q,a,R}$ data points, where $l_{q,a,S}$ and $l_{q,a,R}$ are the number of data
427 points for susceptible and resistant bacterial isolates of pathogen q, respectively. For a data point

428 (x_k, y_k) derived from pathogen q, let $\alpha_k^{a,S}$ and $\alpha_k^{a,R}$ denote the softmax score for the prediction

429 of susceptibility and resistance to antibiotic a, respectively. The prediction sets were decided

430 based on the empirical p-values $p_k^{q,a,S}$ and $p_k^{q,a,R}$, which were calculated according to

431 $$p_k^{q,a,S} = \frac{|i = 1, \ldots, l_{q,a,S}: \tilde{\alpha}_i^{q,a,S} \leq \alpha_k^{a,S}| + 1}{l_{q,a,S} + 1},$$

432 $$p_k^{q,a,R} = \frac{|i = 1, \ldots, l_{q,a,R}: \tilde{\alpha}_i^{q,a,R} \leq \alpha_k^{a,R}| + 1}{l_{q,a,R} + 1},$$

433 where $\tilde{\alpha}_i^{q,a,S}$ and $\tilde{\alpha}_i^{q,a,R}$ denotes the softmax scores calculated using the data points in the

434 calibration dataset for isolates from pathogen q. At a confidence $1 - \varepsilon$, the prediction set was

435 then formed by

436 $$\Gamma_k^{\varepsilon,a} = \{S \text{ if } p_k^{q,a,S} > \varepsilon\} \bigcup \{R \text{ if } p_k^{q,a,R} > \varepsilon\}.$$

437 **Performance**

438 The model's predictive performance was computed based on 257,784, 242,121, and

439 114,410 train; 285,897, 223,661, and 97,542 calibration; and 280,428, 223,665, and 100,927 test

440 individual AST results for *E. coli*, *K. pneumoniae*, and *P. aeruginosa* isolates, respectively. To

441 evaluate the overall model performance, the F1 score was calculated. In addition, the *major error*

442 (ME) rate, defined as the proportion of true susceptible bacterial isolates erroneously predicted as

443 resistant, and the *very major error* (VME) rate, defined as the proportion of true resistant isolates

444 erroneously predicted as susceptible, were also calculated. To measure the performance of the

445 uncertainty control, true predictions were defined as prediction regions containing the true label

446 for each antibiotic, and false predictions were those containing either no labels or only the wrong

447 one.

448 For comparison, the following naive classifier was included (Figure S7). For each of the
449 bacterial species, the frequencies between all pairwise antibiotics and their susceptibility profile
450 were calculated in the training dataset. Then, the ratios $r_{a_R,b} = \frac{|a_R b_R|}{|a_R b_R|+|a_R b_S|}$ and $r_{a_S,b} =$
451 $\frac{|a_S b_R|}{|a_S b_R|+|a_S b_S|}$, between antibiotic a, with resistance (R) and susceptible (S) profiles, and the
452 antibiotic b, with resistance profile, were calculated for all pairwise combinations of antibiotics.
453 Finally, for a new prediction of antibiotic b in y_k, resistance was deemed if the average A =
454 $\sum_{a_j \in x_k} \frac{r_{a_j b}}{n_a}$ was larger than 0.5, where a_j represent the antibiotics in x_k with resistance $j \in$
455 $\{R, S\}$, and n_a is the number of antibiotics in x_k, the isolate was deemed susceptible otherwise.
456 To investigate the effect of patient data, we compared our results with a model based solely on
457 AST results (Figure S6). The VMEs for the three bacterial species were found to be significantly
458 smaller (p-val<0.01) when the original model was compared to the model without patient data
459 and to the naive classifier, as determined by Wilcoxon signed-rank tests.

460

461 **ACKNOWLEDGMENTS**

462 Data from The European Surveillance System – TESSy between 2007 to 2020, provided
463 by Andorra, Albania, Armenia, Austria, Azerbaijan, Bosnia and Herzegovina, Belgium,
464 Bulgaria, Belarus, Switzerland, Cyprus, Czechia, Germany, Denmark, Estonia, Greece, Spain,
465 Finland, France, Georgia, Croatia, Hungary, Ireland, Israel, Iceland, Italy, Kyrgyzstan,
466 Kazakhstan, Liechtenstein, Lithuania, Luxembourg, Latvia, Monaco, Republic of Moldova,
467 Montenegro, Republic of North Macedonia, Malta, Netherlands, Norway, Poland, Portugal,
468 Romania, Serbia, Russian Federation, Sweden, Slovenia, Slovakia, San Marino, Tajikistan,
469 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 model 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)

National Health and Medical Research Council of Australia under the framework of JPI AMR, grant 2031902, SEARCHER

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

491 **SUPPLEMENTAL MATERIALS**

492 Figure S1. The number of susceptible (opaque) and resistant (transparent) bacterial isolates
493 tested for each antibiotic and pathogen in the dataset.

494 Figure S2. Distribution of the number of antibiotics tested per bacterial isolate for each pathogen
495 in the dataset.

496 Figure S3. Distribution of the proportion of bacterial isolates tested against each antibiotic for
497 female (opaque) and male (transparent) patients and for each pathogen across countries in the
498 dataset. The center line, box limits, and whiskers represent the median, upper, and lower
499 quartiles, and 1.5 times the interquartile range, respectively.

500 Figure S4. The proportion of bacterial isolates susceptible to each antibiotic for female (opaque)
501 and male (transparent) patients and for each pathogen across countries in the dataset. The center
502 line, box limits, and whiskers represent the median, upper, and lower quartiles, and 1.5 times the
503 interquartile range, respectively.

504 Figure S5. Histogram of the number of antibiotic susceptibility testing (AST) results used as
505 input for the model during training at each epoch for each pathogen.

506 Figure S6. Performance of the cross-validation. The major error (red) and very major error
507 (VME) rates for each fold are represented by dots, the horizontal line, box limits, and whiskers
508 represent the median, upper, and lower quartiles, and 1.5 times the interquartile range,
509 respectively, across folds.

510 Figure S7. Performance of the model with and without patient data and the naive classifier.

511 Figure S8-12. Empirical and expected error rates and performance of the model after conditional
512 inductive conformal prediction. For each pathogen and penicillins (S8), cephalosporins (S9),
513 quinolones (S10), aminoglycosides (S11), and carbapenems (S12) antibiotics.

514 Figure S13-S15. Performance of the model with different hyperparameters: 1 or 2 heads (1H,
515 2H), 1 or 2 layers (1L, 2L), word embedding of size 64 and position-wise feed-forward layer of
516 128 (small), or word embedding of size 128 and position-wise feed-forward layer of 256 (large),
517 with patient and AST data, or without patient data (no patient), for (S13) *E. coli*, (S14) *K.*
518 *pneumoniae*, and (S15) *P. aeruginosa*.

519 Table S1. Data expansion. For isolate j with A_j AST results, we created a fixed number of data
520 points at each epoch of training, calibration, and testing. The number of AST results, a_k,
521 included in the input sentence x_k of data point k was sampled from a binomial distribution with
522 parameters $N = A_j - 1$ and $p = \frac{D_j}{A_j - 1} - 1$, listed below. If $a_k < 4$, the data point was dismissed.
523 The number of antibiotic susceptibilities to predict for data point k was given by $A_j - a_k$, and
524 the true susceptibility test results are summarized in the sequence y_k. The decision on which
525 antibiotics were included in x_k and y_k was done randomly for all data points.

526 Table S2. Performance of the transformer model. Results from the test dataset: F1-score, major
527 error (ME) rate, and very major error (VME) rate shown for each antibiotic and pathogen,
528 considering all possible numbers of input AST results. Dash (–) indicates combinations of
529 bacterial species and antibiotics that the model does not cover.

530 Table S3. Performance of the transformer model as a function of the number of antibiotic
531 susceptibility testing (AST) results included in the input. Results from the test dataset: F1-score,
532 major error (ME) rate, and very major error (VME) rate shown for each antibiotic, pathogen, and

533 the number of antibiotic susceptibility testing (AST) results included in the input. Dash (–)
534 indicates combinations of bacterial species and antibiotics that the model does not cover. We
535 limited the table to 8 AST results in the input due to the low number of isolates tested for 10 or
536 more antibiotics in the data.

537

538 **TABLES**

539 **Table 1. Bacterial isolates summary.** The included antibiotics, the number of isolates from
540 each species tested for each antibiotic, their resistance rates, and the weights used in the loss
541 function.

Antibiotic	*Escherichia coli*		*Klebsiella pneumoniae*		*Pseudomonas aeruginosa*		Loss function weights
	Isolates	%R[1]	Isolates	%R[1]	Isolates	%R[1]	(S/R)[2]
AMC – Amoxicillin / clavulanic acid	550 937	33%	129 537	42%	–	–	0.45/0.55
AMP – Ampicillin	800 283	57%	–	–	–	–	0.45/0.55
AMX – Amoxicillin	293 369	55%	–	–	–	–	0.45/0.55
PIP – Piperacillin	93 919	50%	21 028	80%	62 218	22%	0.45/0.55
TZP – Piperacillin/ tazobactam	646 892	8%	164 551	29%	152 233	16%	0.15/85
CAZ – Ceftazidime	1 080 891	10%	280 916	30%	159 118	15%	0.3/0.7
CRO – Ceftriaxone	350 947	11%	90 851	30%	–	–	0.3/0.7
CTX – Cefotaxime	990 234	11%	253 971	30%	–	–	0.3/0.7
FEP – Cefepime	336 526	10%	102 213	33%	65 619	16%	0.3/0.7
CIP – Ciprofloxacin	1 117 333	22%	286 899	31%	160 669	20%	0.3/0.7
LVX – Levofloxacin	291 403	23%	73 574	29%	45 636	24%	0.3/0.7
MFX – Moxifloxacin	105 648	26%	23 231	26%	–	–	0.3/0.7
NAL – Nalidixic acid	63 468	27%	16 532	36%	–	–	0.45/0.55
OFX – Ofloxacin	136 704	19%	33 764	36%	–	–	0.3/0.7
AMK – Amikacin	–	–	202 968	9%	130 264	9%	0.15/85
GEN – Gentamicin	1 073 614	9%	278 400	20%	143 645	14%	0.15/85
TOB – Tobramycin	564 295	10%	145 914	27%	119 954	13%	0.15/85
ETP – Ertapenem	–	–	109 997	14%	–	–	0.15/85
IPM – Imipenem	–	–	189 733	10%	121 852	20%	0.15/85

MEM – Meropenem	–	–	242 732	9%	130 352	15%	0.15/85

[1]Proportion of resistance bacterial isolates to each of the antibiotics.

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

Dash (–) indicates combinations of bacterial species and antibiotics that the model does not cover.

FIGURES

Figure 1. The proposed model architecture and uncertainty control. (**A**) The architecture of the proposed model. Both input sequences start with a classification word, *CLS*, followed by the bacterial species name. One of the input sequences also contains patient information (country, gender, age, and sampling date), while the other includes available antibiotic susceptibility testing (AST) data. Both input sequences are fixed to lengths $L = 5$ for the patient information and $L = 21$ for the AST data, padded with the *PAD* word. We use a linear embedding to represent the input words numerically, which are fed into the transformer encoders. The first vectors (*CLS*) from the outputs of each encoder are combined into a linear model and fed to m independent neural networks, each representing one antibiotic. The neural network outputs are two-dimensional vectors indicating susceptibility and resistance, respectively. **(B)** Uncertainty control. The neural network outputs undergo a softmax rescaling. A calibration dataset is used to build empirical distributions of conformity for resistant and susceptible predictions for each pathogen and antibiotic. The prediction regions for the data points k and $k + 1$ are built based on the deviation of the observed softmax score from the empirical distribution and the confidence interval threshold. See Methods for full details.

563 **Figure 2. Performance of the transformer model.** Results from the train and test datasets: (**A**)
564 F1-score, (**B**) major error (ME) rate, and very major error (VME) rate for the transformer model,
565 shown for each antibiotic and pathogen. Note that, due to data restrictions, not all antibiotics are
566 assessed for all three pathogens.

567 **Figure 3. The predictive performance of the model as a function of the number of antibiotic**
568 **susceptibility testing (AST) results included in the input.** Major error (ME) and very major
569 error (VME) rates from light to dark: 4 to 8 AST results are shown for each antibiotic and
570 pathogen. Note that, due to data restrictions, not all antibiotics are assessed for all three
571 pathogens.

572 **Figure 4. The proportion of correct predictions as a function of the number of labels in the**
573 **prediction set for *E. coli* isolates.** The proportion of correct predictions with a single label
574 (light) and multiple labels (dark), major errors (MEs), and very major errors (VMEs) with a
575 single label (light) and empty set (dark) predictions for resistant (R) and susceptible (S) isolates,
576 shown for each antibiotic and pathogen using three different confidence levels: 90%, 95%, and
577 97.5%. **(A)** The proportions are shown for each antibiotic at three confidence levels (90%, 95%,
578 and 97.5%), considering all possible numbers of input AST results. **(B)** The proportions are
579 shown as a function of the number of input AST results (90% confidence level).

580 **Figure 5. The proportion of correct predictions as a function of the number of labels in the**
581 **prediction set for *K. pneumoniae* isolates.** The proportion of correct predictions with a single
582 label (light) and multiple labels (dark), major errors (MEs), and very major errors (VMEs) with a
583 single label (light) and empty set (dark) predictions for resistant (R) and susceptible (S) isolates,
584 shown for each antibiotic and pathogen using three different confidence levels: 90%, 95%, and
585 97.5%. **(A)** The proportions are shown for each antibiotic at three confidence levels (90%, 95%,

586 and 97.5%), considering all possible numbers of input AST results. **(B)** The proportions are
587 shown as a function of the number of input AST results (90% confidence level).

588 **Figure 6. The proportion of correct predictions with respect to the number of labels in the**
589 **prediction set for *P. aeruginosa* isolates.** The proportion of correct predictions with a single
590 label (light) and multiple labels (dark), major errors (MEs), and very major errors (VMEs) with a
591 single label (light) and empty set (dark) predictions for resistant (R) and susceptible (S) isolates,
592 shown for each antibiotic and pathogen using three different confidence levels: 90%, 95%, and
593 97.5%. **(A)** The proportions are shown for each antibiotic at three confidence levels (90%, 95%,
594 and 97.5%), considering all possible numbers of input AST results. **(B)** The proportions are
595 shown as a function of the number of input AST results (90% confidence level).

596

597 **REFERENCES**

598 1. World Health Organization. 2015. Global action plan on antimicrobial resistance.

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

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

604 4. Mercer DK, Torres MDT, Duay SS, Lovie E, Simpson L, von Köckritz-Blickwede M, de la
605 Fuente-Nunez C, O'Neil DA, Angeles-Boza AM. 2020. Antimicrobial Susceptibility

606 Testing of Antimicrobial Peptides to Better Predict Efficacy. Frontiers in Cellular and
607 Infection Microbiology 10:326.

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

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

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

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

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

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

626 11. Dolejska M, Papagiannitsis CC. 2018. Plasmid-mediated resistance is going wild. Plasmid
627 99:99–111.

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

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

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

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

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

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

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

649 19. Yu K-H, Beam AL, Kohane IS. 2018. Artificial intelligence in healthcare. Nature
650 biomedical engineering 2:719–731.

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

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

655 22. Ali T, Ahmed S, Aslam M. 2023. Artificial Intelligence for Antimicrobial Resistance
656 Prediction: Challenges and Opportunities towards Practical Implementation. Antibiotics
657 (Basel) 12:523.

658 23. Sakagianni A, Koufopoulou C, Feretzakis G, Kalles D, Verykios VS, Myrianthefs P,
659 Fildisis G. 2023. Using Machine Learning to Predict Antimicrobial Resistance—A
660 Literature Review. Antibiotics 12:452.

661 24. The Medical Futurist. 2022. FDA-approved A.I.-based algorithms.
662 https://medicalfuturist.com/fda-approved-ai-based-algorithms/

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

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

667 27. Feretzakis G, Loupelis E, Sakagianni A, Kalles D, Martsoukou M, Lada M, Skarmoutsou
668 N, Christopoulos C, Valakis K, Velentza A, others. 2020. Using machine learning
669 techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a
670 general hospital in Greece. Antibiotics 9:50.

671 28. Feretzakis G, Sakagianni A, Loupelis E, Kalles D, Skarmoutsou N, Martsoukou M,
672 Christopoulos C, Lada M, Petropoulou S, Velentza A, others. 2021. Machine learning for
673 antibiotic resistance prediction: A prototype using off-the-shelf techniques and entry-level
674 data to guide empiric antimicrobial therapy. Healthcare informatics research 27:214–221.

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

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

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

683 32. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K,
684 Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A,
685 Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E,

686 Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals
687 O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein
688 structure prediction with AlphaFold. Nature 596:583–589.

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

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

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

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

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

702 38. Ahmad M, Khan AU. 2019. Global economic impact of antibiotic resistance: A review.
703 Journal of global antimicrobial resistance 19:313–316.

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

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

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

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

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

721

722

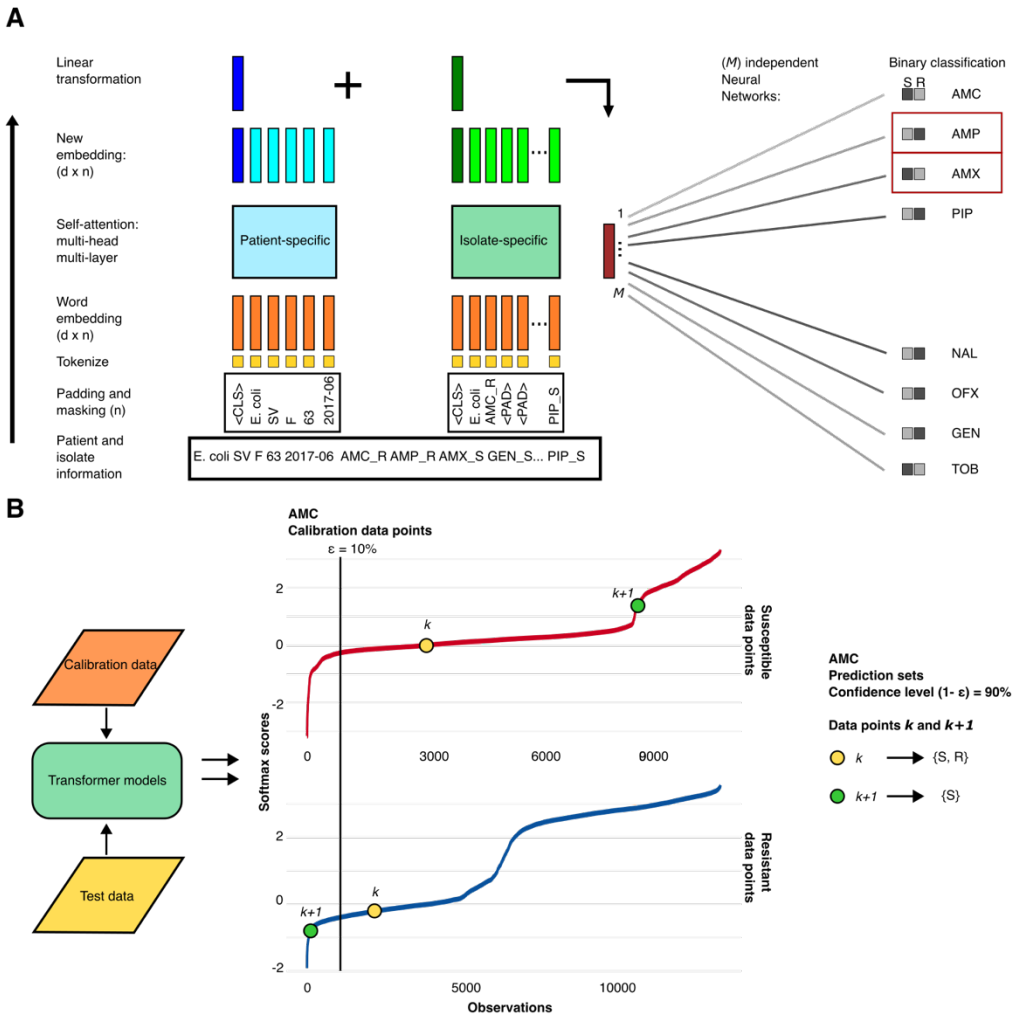

Figure 1. The proposed model architecture and uncertainty control. (A) The architecture of the proposed model. Both input sequences start with a classification word, *CLS*, followed by the bacterial species name. One of the input sequences also contains patient information (country, gender, age, and sampling date), while the other includes available antibiotic susceptibility testing (AST) data. Both input sequences are fixed to lengths L=5 for the patient information and L=21 for the AST data, padded with the *PAD* word. We use a linear embedding to represent the input words numerically, which are fed into the transformer encoders. The first vectors (*CLS*) from the outputs of each encoder are combined into a linear model and fed to *m* independent neural networks, each representing one antibiotic. The neural network outputs are two-dimensional vectors indicating susceptibility and resistance, respectively. **(B)** Uncertainty control. The neural network outputs undergo a softmax rescaling. A calibration dataset is used to build empirical distributions of conformity for resistant and susceptible predictions for each pathogen and antibiotic. The prediction regions for the data points k and $k+1$ are built based on the deviation of the observed softmax score from the empirical distribution and the confidence interval threshold. See Methods for full details.

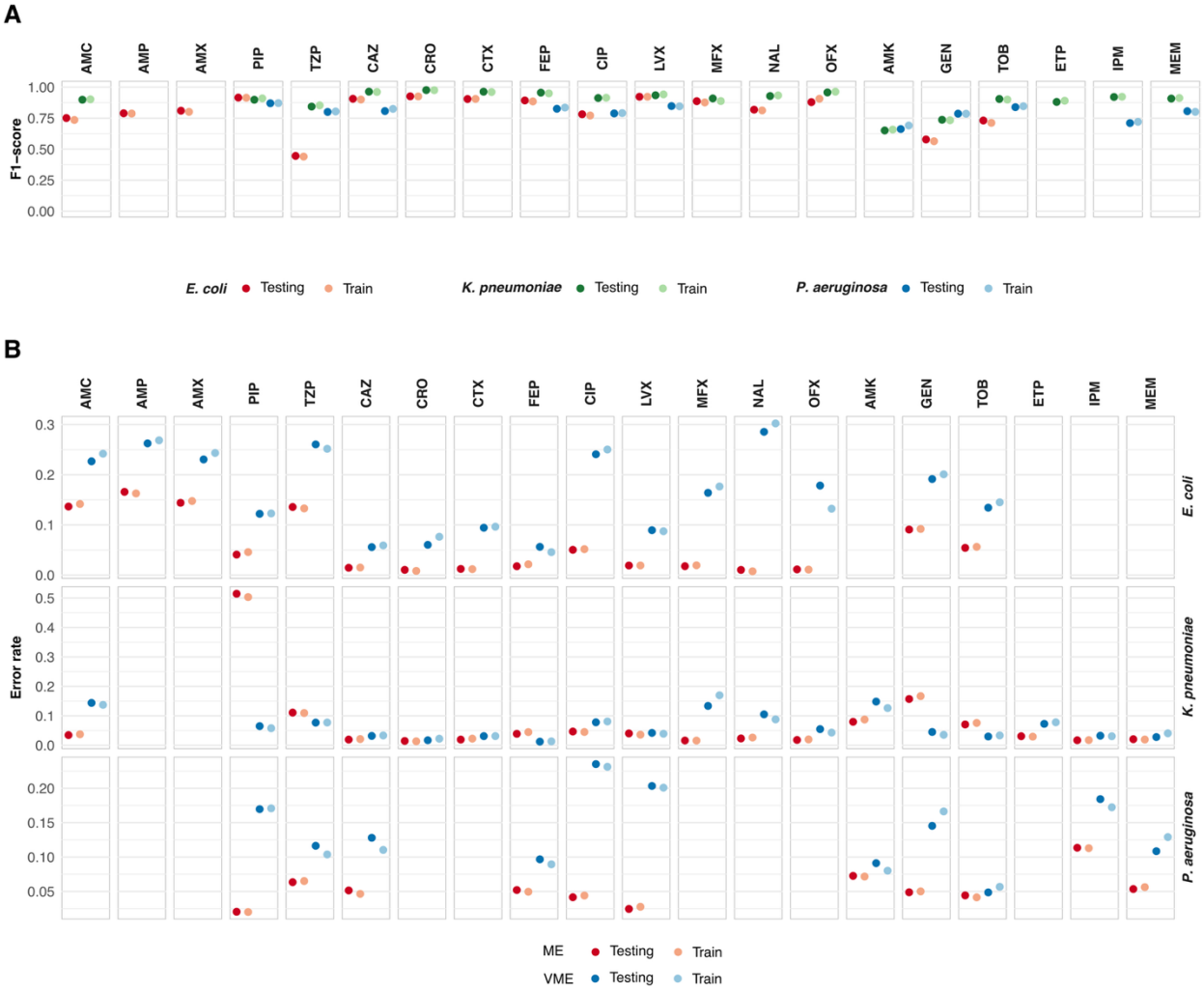

Fig. 2. Performance of the transformer model. Results from the train and testing datasets: **(A)** F1-score, **(B)** major error (ME) rate, and very major error (VME) rate for the transformer model shown for each antibiotic and pathogen. Note that, due to data restrictions, not all antibiotics are assessed for all three pathogens.

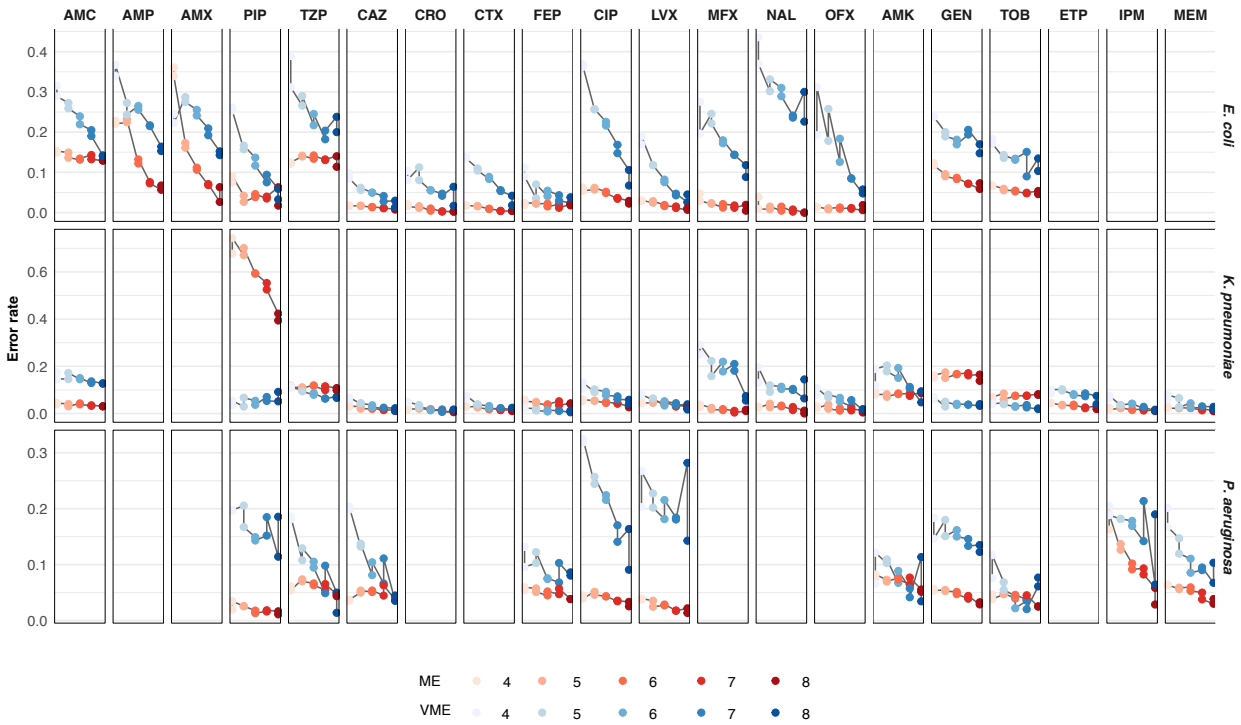

Fig. 3. The predictive performance of the model as a function of the number of antibiotic susceptibility testing (AST) results included in the input. Major error (ME) and very major error (VME) rates from light to dark: 4 to 8 AST results are shown for each antibiotic and pathogen. Note that, due to data restrictions, not all antibiotics are assessed for all three pathogens.

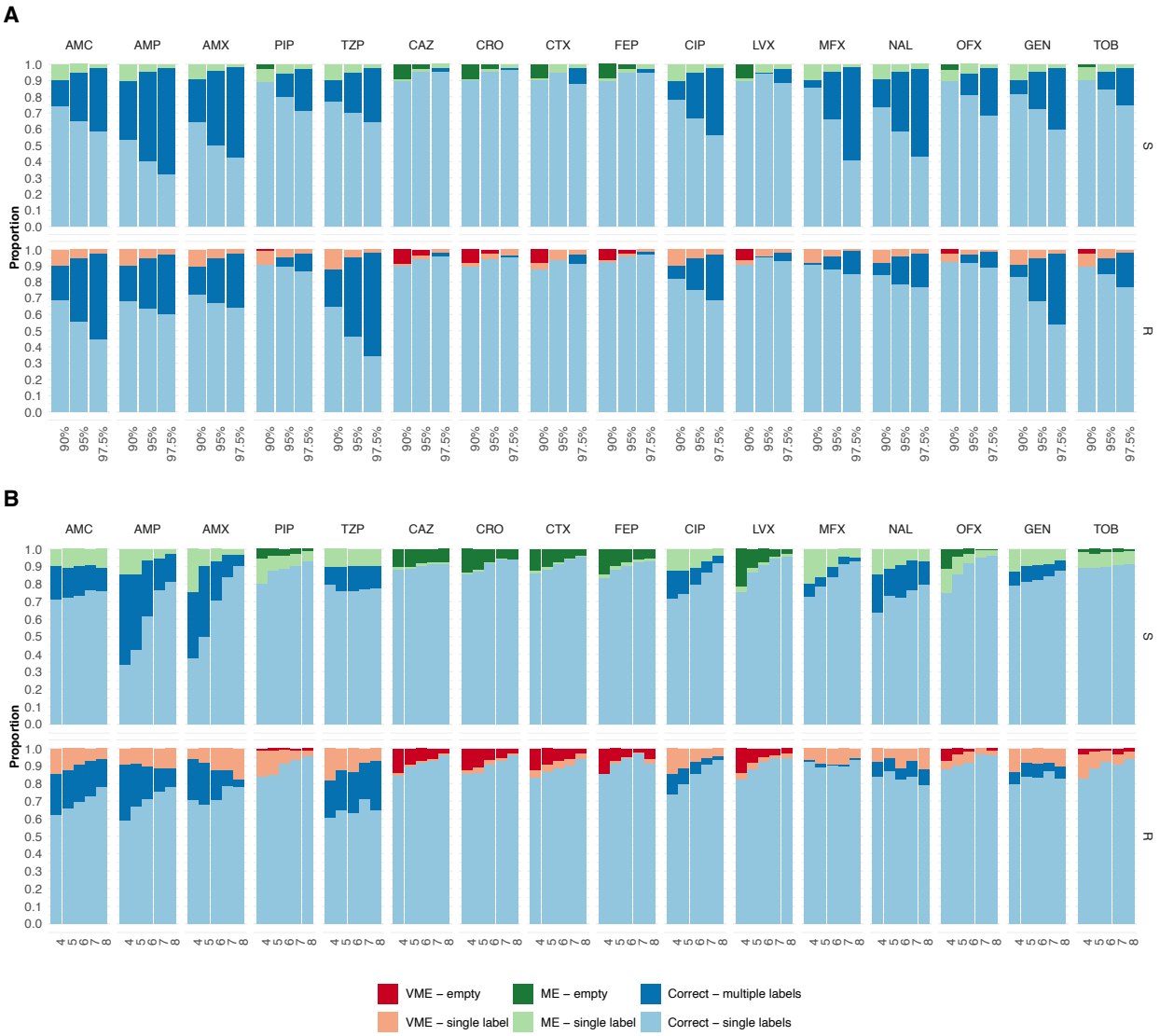

Fig. 4. The proportion of correct predictions with respect to the number of labels in the prediction set for *E. coli* isolates. The proportion of correct predictions with a single label (light) and multiple labels (dark), major errors (MEs), and very major errors (VMEs) with a single label (light) and empty set (dark) predictions for resistant (R) and susceptible (S) isolates, shown for each antibiotic and pathogen using three different confidence levels: 90%, 95%, and 97.5%. **(A)** The proportions are shown for each antibiotic at three confidence levels (90%, 95%, and 97.5%), considering all possible numbers of input AST results. **(B)** The proportions are shown as a function of the number of input AST results (90% confidence level).

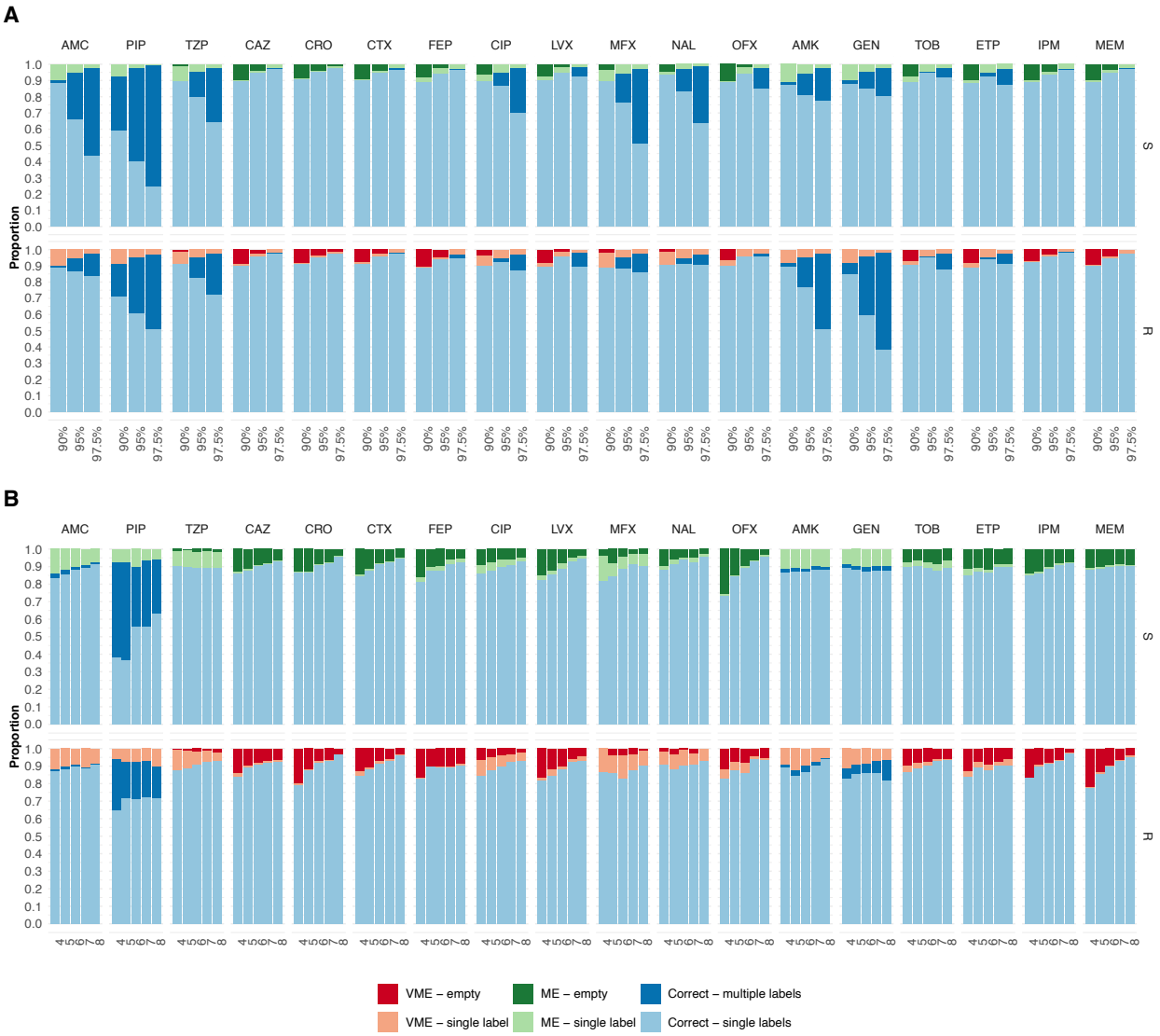

Fig. 5. The proportion of correct predictions with respect to the number of labels in the prediction set for *K. pneumoniae* isolates. The proportion of correct predictions with a single label (light) and multiple labels (dark), major errors (MEs), and very major errors (VMEs) with a single label (light)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

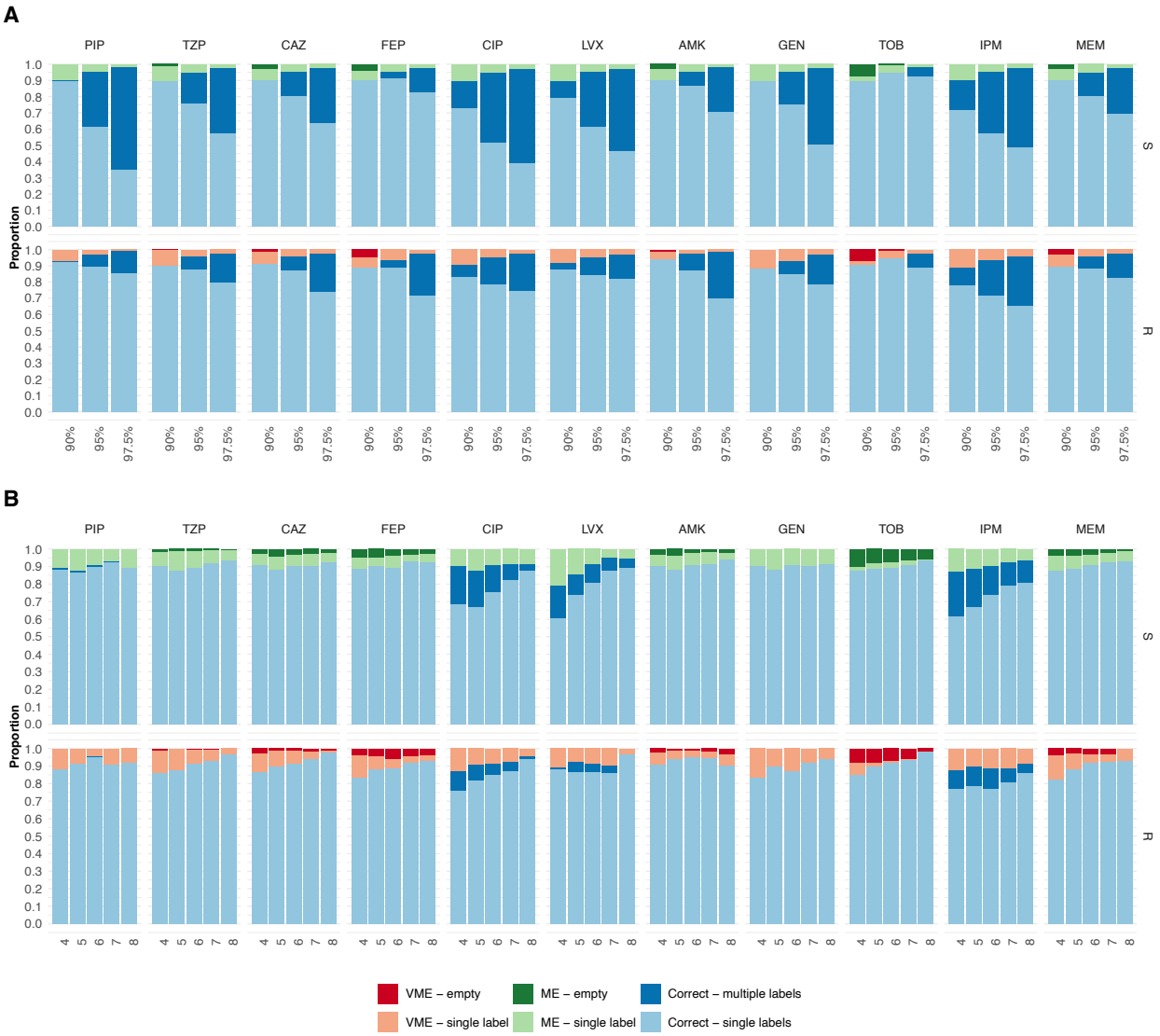

Fig. 6. The proportion of correct predictions with respect to the number of labels in the prediction set for *P. aeruginosa* isolates. The proportion of correct predictions with a single label (light) and multiple labels (dark), major errors (MEs), and very major errors (VMEs) with a single label (light) and empty set (dark) predictions for resistant (R) and susceptible (S) isolates, shown for each antibiotic and pathogen using three different confidence levels: 90%, 95%, and 97.5%. **(A)** The proportions are shown for each antibiotic at three confidence levels (90%, 95%, and 97.5%), considering all possible numbers of input AST results. **(B)** The proportions are shown as a function of the number of input AST results (90% confidence level).