Direct Antimicrobial Resistance Prediction from MALDI-TOF mass spectra profile in clinical isolates through Machine Learning

Caroline Weis (1) (2)*, Aline Cuénod (3) (4), Bastian Rieck (1) (2), Felipe Llinares-López (1) (2), Olivier Dubuis (8), Susanne Graf (7), Claudia Lang (8), Michael Oberle (9), Kirstine K. Soegaard (3) (4), Michael Osthoff (5) (6), Karsten Borgwardt (1) (2)*+, Adrian Egli (3) (4)*+

(1) Department of Biosystems and Engineering, ETH Zürich, Basel, Switzerland
(2) SIB Swiss Institute of Bioinformatics, Switzerland
(3) Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland
(4) Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland
(5) Division of Infectious Diseases and Hospital Epidemiology, University Hospital and University of Basel, Basel, Switzerland
(6) Department of Internal Medicine, University Hospital and University of Basel, Basel, Switzerland
(7) Department for Microbiology, Cantonal Hospital Baselland, Liestal, Switzerland
(8) Clinical Microbiology, Viollier AG, Allschwil, Switzerland
(9) Institute for Laboratory Medicine, Medical Microbiology, Cantonal Hospital of Aarau, Aarau, Switzerland

+ equally contributed
* corresponding:
Caroline Weis
Machine Learning and Computational Biology Lab
Mattenstrasse 26
4058 Basel, Switzerland
Email: caroline.weis@bsse.ethz.ch
Phone: +41 61 387 34 39

Prof. Dr. Karsten Borgwardt
Machine Learning and Computational Biology Lab
Mattenstrasse 26
4058 Basel, Switzerland
Email: karsten.borgwardt@bsse.ethz.ch
Phone: +41 61 387 34 20

Prof. Dr. Dr. Adrian Egli, MD PhD
Clinical Bacteriology and Mycology
University Hospital Basel
Petersgraben 4
4031 Basel, Switzerland
Email: adrian.egli@usb.ch
Phone: +41 61 556 57 49

Abstract

Early administration of effective antimicrobial treatments improves the outcome of infections. Culture-based antimicrobial resistance testing allows for tailored treatments, but takes up to 96h. We present a revolutionary approach to predict resistance with unmatched speed within 24h, using calibrated logistic regression and LightGBM-classifiers trained on species-specific MALDI-TOF mass spectrometry measurements. For this analysis, we created an unprecedented large, publicly-available dataset combining mass spectra and resistance information. Our models provide highly valuable treatment guidance 12–72h earlier than classical approaches. Rejection of uncertain predictions enables quality control and clinically-applicable sensitivities and specificities for the priority pathogens *Staphylococcus aureus*, *Escherichia coli*, and *Klebsiella pneumoniae*.

Main

Antimicrobial resistant bacteria and fungi pose an important and increasing threat at global levels[1,2]. Infections with antimicrobial resistant pathogens are associated with substantial morbidity, mortality, and healthcare costs[3]. Effective and immediate antimicrobial treatment improves the outcome of an infection[4,5]. Foremost, the empiric antimicrobial therapy and dosage should cover resistance profiles of presumed pathogens, and also consider host-specific factors such as age, kidney function, and previous patient history. Early identification of the pathogen already impacts the antimicrobial treatment[6,7], however, only a detailed resistance profile permits to fully optimize treatments. With current culture-based methods, the time from sample collection to susceptibility reporting takes up to 96h[8,9]. In the meantime, empiric treatment usually consists of broad spectrum antimicrobials until results from phenotypic antimicrobial resistance testing are available. Moreover, rapidly available results would benefit all goals of antibiotic stewardship[10], which include providing effective treatments, reducing unnecessary antibiotic use, and fighting the development of antibiotic resistance. Thus, the rapid assessment of antibiotic resistance is of utmost importance for global health efforts.

Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry (MS) characterises the mass spectral composition of single bacterial or fungal colonies within *minutes*[11–13]. Single colonies are usually available within 24h after sample collection. These mass spectra enable precise and low-cost microbial identification and the MALDI-TOF MS technology is nowadays the most commonly used method for identification at species levels[7,14,15]. However, the potential of extracting additional information directly from acquired MALDI-TOF mass spectra data (mass spectra) is largely unexplored. The possibility to infer antimicrobial resistance has just recently been introduced e.g. via detection of specific marker masses associated with selected resistance mechanisms[16]. To date, the list of identified marker masses remains very limited[11]. To the best of our knowledge, no study predicted clinically-applicable antimicrobial resistance across multiple antimicrobials and species. Our study aims to harness the full potential of MALDI-TOF MS and predict antimicrobial resistance through machine learning methods. In this context, machine learning efforts are rare[17,18] and stymied by the lack of large, publicly-available, high-quality benchmark datasets[19,20]. Furthermore, no study reports confidence estimates for their antimicrobial resistance predictions or analyse stability over time or validations with multiple sample collection sites.

We present the first study aiming for clinically-applicable mass spectra based antimicrobial resistance prediction. As a foundation, we have curated the **D**atabase of **R**es**I**stance against **A**ntimicrobials with **M**ALDI-TOF mass **S**pectrometry (*DRIAMS)*: an unparalleled, routinely-acquired number of bacterial and fungal MALDI-TOF mass spectra and corresponding antimicrobial resistance profiles, addressing the current lack of a large-scale, high-quality, and real-world dataset. Using this foundation, we developed an antimicrobial resistance prediction approach. We investigate its efficacy for a broad range of bacterial and fungal pathogens and showcase the predictive power on three clinical most relevant scenarios with *Staphylococcus aureus, Escherichia coli*, and *Klebsiella pneumoniae*, respectively, and observe superior performance of logistic regression and LightGBM models compared to other algorithms. Our calibrated classifiers with built-in support for sample rejection provide safe predictions for patients, while reaching sensitivity and specificity values larger than 90% for oxacillin in *S. aureus* and ceftriaxone in *E. coli* and *K. pneumoniae*. As mass spectra are available after overnight culture already, our approach provides guidance for early antimicrobial patient treatment decisions, much sooner than classical culture-based phenotypic testing. We discuss the limitations of our approach through temporal and across-site validation and finally show that it can be successfully replicated, when retraining on data from other healthcare centers. We make *DRIAMS* publicly available to serve as a benchmark for future research in this new field of machine learning using MALDI-TOF mass spectra. All results are reproducible with code available on GitHub at https://github.com/BorgwardtLab/maldi_amr.

Results

DRIAMS: Clinical routine database combining MALDI-TOF mass spectra and antimicrobial resistance profiles

We have collected an unprecedented dataset of MALDI-TOF mass spectra and resistance profile data from four different institutions, which we have termed *DRIAMS*. The raw dataset collection comprises a total of 303'195 mass spectra and 768'300 antimicrobial resistance labels, pre-processed and organized in four subcollections (*DRIAMS-A* to *-D*) from four institutions in Switzerland. *DRIAMS-A,* the largest collection, was collected at the University Hospital Basel (Switzerland) and is used for the main analysis. A summary of the 71 antimicrobials included in *DRIAMS-A*, including the number of spectra and antimicrobial resistance ratio, is given in **Suppl. Tab. 1** and **Suppl. Tab. 2**.

Machine learning for MALDI-TOF MS based antimicrobial susceptibility prediction

For all analyses, we define mass spectra based antimicrobial resistance prediction as a binary classification scenario. To this end, we preprocessed and binned mass spectra measurement points from 2'000 Da to 20'000 Da into a fixed bin size of 3 Da, in order to obtain fixed-length feature vectors for each sample. Next, we converted the antimicrobial resistance categories, which are either given as susceptible, intermediate or resistant classes, into a binary label (susceptible vs. intermediate/resistant). A positive label corresponds to either intermediate or resistant samples, while a negative label is given susceptible samples (in most of the scenarios we consider, the positive class will be the minority class). We trained multiple classifiers, i.e. logistic regression (LR), random forests (RF), support vector machines with different kernels (SVM), and gradient-boosted decision trees (LightGBM), to predict resistance to each individual antimicrobial. Overall, we observe that LR, RF, and LightGBM are among the best-performing classifiers. Since RF tends to exhibit a lower generalisation performance (see the Methods section for more details), we will focus on LR and LightGBM in the main text and defer an in-depth and critical comparison between all classifiers to the Supplemental Materials. In the following, we report the common machine learning metrics 'area under the receiver operator characteristic' (AUROC) and 'area under the precision-recall curve' (AUPRC), the latter of which is specifically suited given the class imbalance. **Fig. 1** depicts the workflow from data collection, data processing, and prediction results.

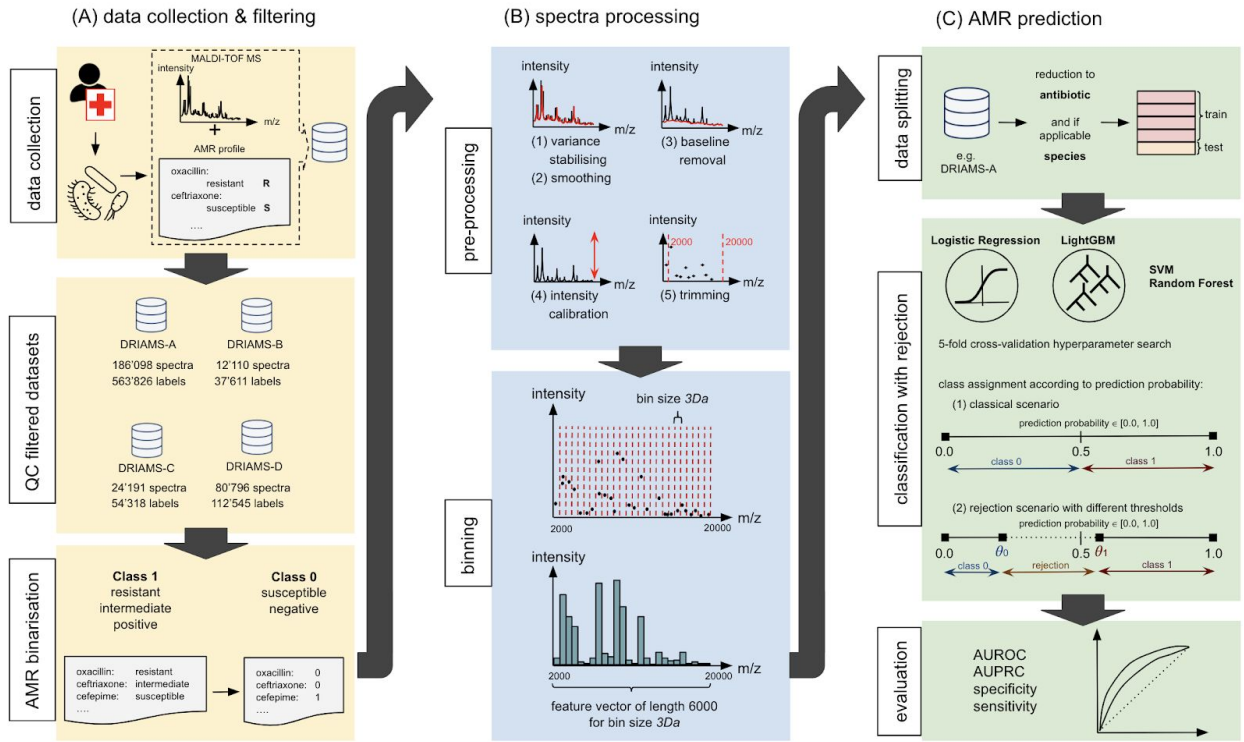


Figure 1: MALDI-TOF MS based AMR prediction workflow. Workflow of MALDI-TOF MS data preprocessing and antibiotic resistance prediction using machine learning. **A.** (i) Data collection: Samples are taken from infected patients, pathogens are cultured, and their mass spectra and resistance profiles are determined. Spectra and resistance are extracted from the MALDI-TOF MS and laboratory information system, corresponding entries matched and added to a dataset. (ii) Quality control (QC) filtered datasets: After several steps ensuring quality control the datasets are added to *DRIAMS*. (iii) Antimicrobial resistance (AMR) binarisation: antimicrobial resistance is defined as a binary classification scenario, with the positive class represented by all labels leading to the antimicrobial not being administered, i.e. intermediate or resistant, and positive, while the negative class represents susceptible or negative samples. **B.** (i) Pre-processing: Cleaning of mass spectra. (ii) Binning: Binning spectra into equal-sized feature vectors for machine learning. **C.** (i) Data splitting: Data is split into 80% training and 20% test, stratified by both class and species. (ii) Classification with rejection: Antimicrobial resistance classifiers are trained using a 5-fold cross-validation for hyperparameter search, using classification algorithms logistic regression, LightGBM, SVM (with linear and rbf kernel), and random forests. Classes are either assigned in a classical fashion, by assigning the positive class to samples with a positive class score larger than 0.5, or with rejection, by refusing to assign classes to samples between two score thresholds θ_0 and θ_1. (iii) Evaluation: Predictive performance is measured in metrics commonly used in machine learning, such as AUROC and AUPRC, and the medical community, such as specificity and sensitivity.

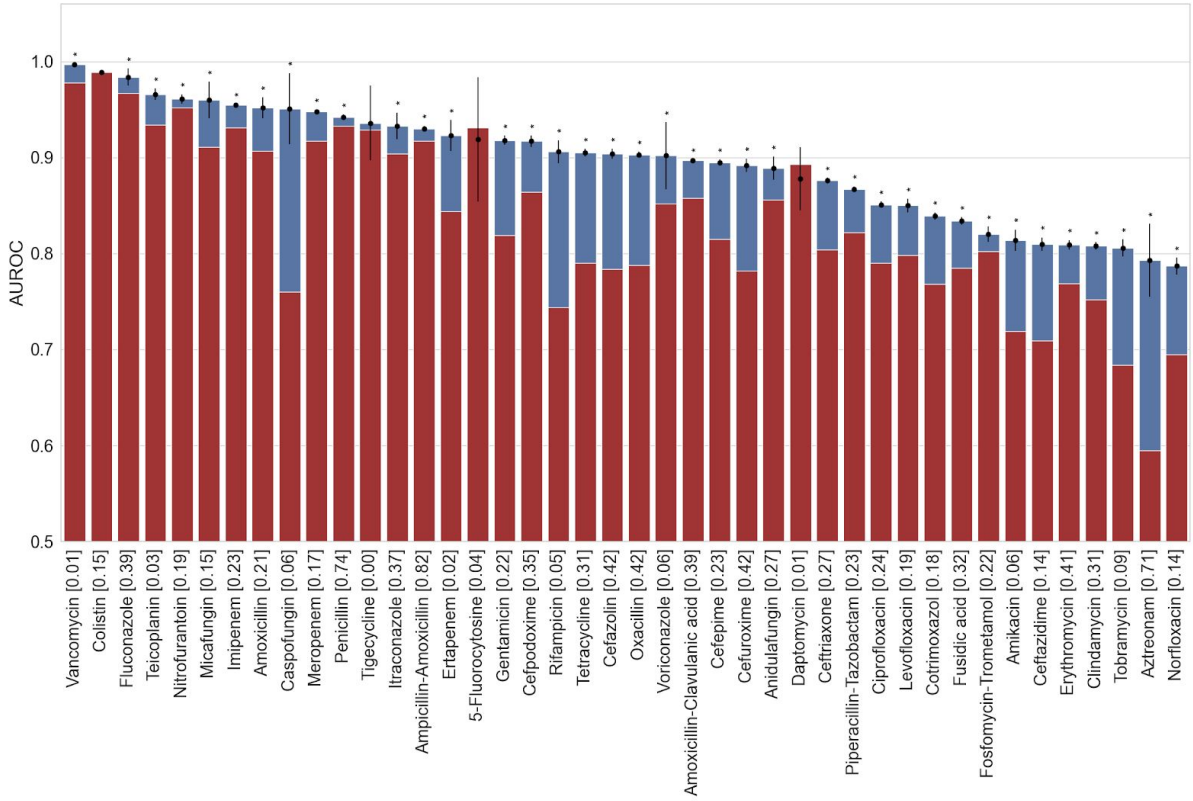


Figure 2: Improved antimicrobial resistance prediction based on MALDI-TOF mass spectra combining all species compared to species information alone. AUROC values of logistic regression classifiers trained on data combining all samples with labels available for each antimicrobial prediction task in *DRIAMS-A*. The blue bars depict predictive performance using spectral data as features. The red bars show the predictive performance when using species label information only. The fractions of resistant/intermediate samples in the training data are indicated in brackets after the antibiotic name. Reported metrics and error bars are the mean and standard deviation of 10 repetitions with different random train-test-splits. We see improvement in predictive performance in 39 out of 42 antimicrobials, with statistically significant differences indicated by stars. In the current clinical procedure, at the time of MALDI-TOF measurement, treatment decisions are performed based on species information alone.

As an initial experiment, we want to evaluate whether prediction performance using our clinical routine mass spectra dataset combining all measured species is superior, compared to inferring resistance from knowing only microbial species. This information is crucial, as the current procedure to modify antibiotic treatments at the time of MALDI-TOF MS is a decision informed strongly by the identified species. A predictive performance from mass spectra higher than from species information alone is evidence that a mass spectra based antimicrobial resistance prediction increases the information available for clinical decision-making, at a much earlier time point than phenotypic testing is available (in median 22h earlier, **Suppl. Fig. 5**). This analysis was conducted on the *DRIAMS-A* dataset, as it is most comprehensive in terms of sample size and collection time frame . An overall assessment depicted in **Fig. 2** shows AUROC predictive performance for the 42 studied antibiotics, which had enough samples from each class to apply machine learning. For 23 out of the 42 antibiotics, we reached AUROC values above 0.90. Moreover, for 39 antibiotics, we observe that the improvements in performance gained from using mass spectra are statistically significant as opposed to merely using species information for resistance prediction . The results clearly demonstrate the predictive power of mass spectra based antimicrobial resistance prediction, and their ability to provide more information than species information alone. Next, we narrow our focus to three pathogens frequently causing infections in clinical practice and assess the performance of a classifier based on specific species–antibiotic combinations.

Species-specific AMR prediction yields high performance for clinically-relevant pathogens

In clinical practice, physicians regularly adapt the antimicrobial treatment for a specific bacterial species. For detailed analysis, we select three species *S. aureus, E. coli*, and *K. pneumoniae*; all of which have been named 'priority pathogens' by the World Health Organization[21]. We first analyse to what extent a logistic regression model is capable of predicting resistance for specific bacterial species (see **Fig. 3**), observing high overall performance in both AUROC and AUPRC curves. Additionally, we describe each plot in terms of specificity and sensitivity to permit comparability with medical literature. All depicted antibiotics were selected according to their clinical relevance, with respect to the choice of treatment or for their utility in screening for resistant pathogens e.g. for hospital epidemiology.

In the following, we exemplarily describe the scenario for *S. aureus*. The prediction of oxacillin resistance reaches a high performance with AUROC of 0.90 and AUPRC of 0.80 at a positive (i.e. resistant or intermediate) class ratio of 19.2%. According to laboratory protocols used in *DRIAMS-A*, for *S. aureus* strains, the reported susceptibility of nine beta-lactam antibiotics are inferred from the oxacillin susceptibility test results. Furthermore, oxacillin resistance detects methicillin-resistant *S. aureus* (MRSA) strains[22]. Generation of such early resistance information is expected to have a significant impact in treatment adaptations and patient management.

Overall, this experiment demonstrates that a classifier can achieve useful prediction performance on this task. We initially selected logistic regression (LR) because of its scalability and interpretability, and we observe that it exhibits high performance in *all* scenarios; in fact, its performance is on a par with more complex classifiers, such as LightGBM. Please refer to the Methods section and **Suppl. Fig. 1** for a more detailed comparison with other classifiers. Subsequently, we will show results for LR and LightGBM in the main text, with LR representing highly-scalable and interpretable classifiers and LightGBM representing more expressive (but also more complex) ones.

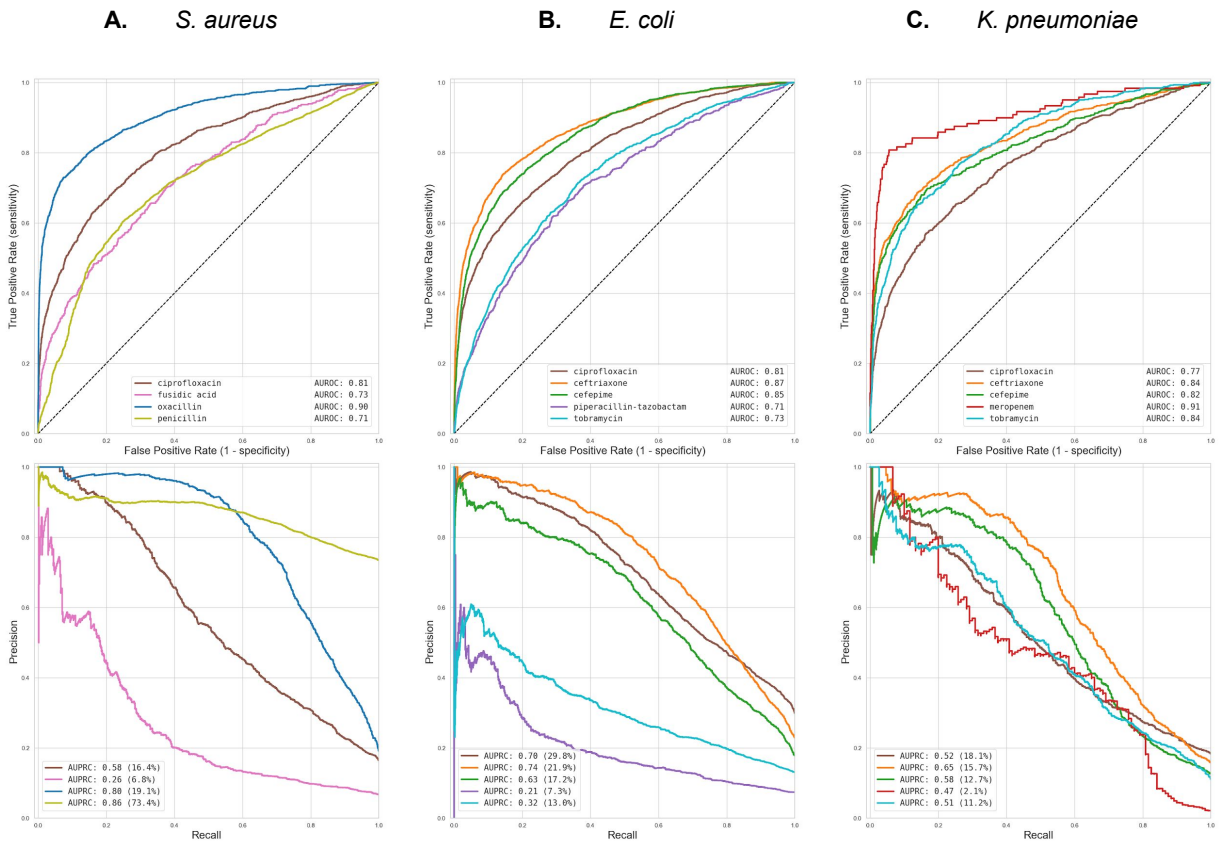


Figure 3: AUROC and AUPRC curves for antimicrobial resistance prediction using logistic regression. For each antibiotic a *DRIAMS-A* subset was created, which was further divided into stratified training and testing data according to species prevalence. We train a logistic regression predictor for the clinically relevant species *S. aureus, E. coli*, and *K. pneumoniae*, paired with several antibiotics which are considered for infections with these species. **A.** For *S. aureus*, our model performs best for oxacillin, with an AUROC of 0.90. This is particularly relevant, as for *S. aureus*, the resistance to nine other beta-lactam antibiotics (including amoxicillin/clavulanic acid and ceftriaxone) is directly derived from oxacillin resistance. **B.** For *E. coli*, the best-performing predictor is that for ceftriaxone, followed by that for cefepime - both critical antibiotics indicating an extended beta-lactamase (ESBL) if resistant. **C.** Finally, for *K. pneumoniae*, meropenem exhibits the highest performance of 0.91 AUROC. Compared to the other scenarios, its ROC curve has a larger step size, but with over 500 test samples, the sample size is comparable to the other antibiotics. The next best performance is achieved by ceftriaxone and tobramycin (both with AUROC 0.84). Despite having the same AUROC, we observe markedly different behaviours in their precision–recall curves. The performances using random forests, support vector machines with linear and rbf kernel, and gradient-boosted decision trees can be found in **Suppl. Fig. 1**.

species	antibiotic	model	θ_0 (keep scores less than and assign class 0)	θ_1 (keep scores larger than and assign class 1)	percentage rejected samples	specificity	sensitivity	TP	FP	TN	FN
E.coli	ceftriaxone	LightGBM	0.025	0.5	72	0.87	0.99	1273	194	1311	13
			0.05	0.5	56	0.94	0.96	1273	194	2830	52
			0.075	0.5	43	0.95	0.91	1273	194	4048	121
			0.1	0.5	34	0.96	0.86	1273	194	4896	210
			0.125	0.5	26	0.97	0.81	1273	194	5601	291
			0.15	0.5	21	0.97	0.78	1273	194	6065	357
S.aureus	oxacillin	LightGBM	0.025	0.5	65	0.96	0.98	881	73	1708	18
			0.05	0.5	44	0.98	0.93	881	73	3219	66
			0.075	0.5	32	0.98	0.87	881	73	4073	127
			0.1	0.5	24	0.98	0.83	881	73	4601	176
			0.125	0.5	19	0.99	0.8	881	73	4935	221
			0.15	0.5	16	0.99	0.77	881	73	5175	268
K. pneumoniae	ceftriaxone	LightGBM	0.025	0.5	86	0.88	0.98	382	49	370	9
			0.05	0.5	66	0.97	0.88	382	49	1483	51
			0.075	0.5	44	0.98	0.74	382	49	2657	132
			0.1	0.5	31	0.99	0.67	382	49	3319	187
			0.125	0.5	22	0.99	0.61	382	49	3803	241
			0.15	0.5	16	0.99	0.58	382	49	4114	276

Table 1: Rejection scenarios of calibrated LightGBM classifiers. *DRIAMS-A* trained classifiers were calibrated such that their classification scores become probabilities. We simulate scenarios in which samples are rejected if their classification probability falls below a certain threshold. The stricter such a threshold, the more samples are being rejected such that only samples with a high predictive probability are being kept. This is seen as a 'trade-off' analysis, where large thresholds lead to better prediction for *some* samples, at the expense of a higher rate of samples that are rejected altogether, i.e. not classified at all. The table illustrates the trade-off between different rejection thresholds for both classes and the amount of rejected/unclassified samples, and how it affects specificity and sensitivity performance. Setting the rejection threshold for the negative (susceptible) class very strictly leads to high performance in sensitivity; the strictest thresholds cause a high percentage of rejected samples and a drop in specificity performance. However a threshold exists that keeps both specificity and sensitivity performing well. The other columns depict a class rejection threshold of 0.5, which equals no rejection of positive classified samples. The classifiers are very accurate in positive class prediction and setting a rejection threshold for positive samples does not improve predictive performance drastically. Results with a logistic regression model can be found in **Suppl. Tab. 3.**

Calibrated classifiers allow for accurate probability estimates

Having analysed the performance of a simple classifier in certain scenarios, we want to focus on improving predictions in order to ensure high-quality results that maximise patient safety. To this end, we *calibrate* our classifiers, i.e. we change the 'scores' a classifier assigns to a sample to actual probabilities[23]. This *calibration* procedure makes all prediction probabilities directly interpretable and permits us to simulate different scenarios, in which we let classifiers reject samples below a certain probability score. This has the effect of keeping only those samples for which a classifier can make a confident prediction , whereas samples that cannot be readily classified are discarded. A calibrated classifier therefore has the option to 'refuse' to classify a sample. This decreases the number of classifiable samples but also increases the confidence for a clinician in those isolates which have been classified. In particular, it prevents a model from reporting predictions that are likely to be wrong, which reduces diagnostic performance and may result in non-effective treatments. The relationships between rejection threshold and accuracy, specificity, and sensitivity are depicted in **Suppl. Fig. 2**.

Rejection option permits quality improvement

For the clinical utility of a model, it should be able to predict resistant samples with high probability, because a false negative (false susceptible) means that patients are treated with an ineffective antibiotic. As mentioned above, this requires a calibrated classifier, whose prediction scores are probabilities. Thus, scores close to 0.50 indicate the most uncertain predictions. We have designed an *asymmetric* rejection scenario using two thresholds $\theta_0 \in [0.0,\ 0.5]$ and $\theta_1 \in [0.5, 1.0]$ to handle predictions: if the predicted score of a sample exceeds θ_1, we predict the positive class, whereas we predict the negative class if the score is below θ_0. Samples with a predicted probability between θ_0 and θ_1 will be discarded, meaning that we require a certain *minimum confidence* for both classes in order for a classifier to perform a prediction, and we make it possible to *refuse* a prediction entirely. By default, $\theta_0 = \theta_1 = 0.5$, meaning that no samples will be rejected.

The values of specificity and sensitivity are connected by their definition in terms of contingency table values (see the Methods section for more details); putting a stricter threshold on the negative class, for instance, will *increase* sensitivity but *decrease* specificity. Since we already observe a specificity of 99% (see **Suppl. Tab. 4**) even without adjusting θ_1 (i.e. there is no need for us to reject positively-labelled examples), we subsequently leave θ_1 fixed, unless mentioned otherwise. By contrast, decreasing θ_0 results in increasing the percentage of rejected samples.

Improved diagnostic performance of antimicrobial resistance prediction with rejection option

Tab. 1 shows the results of calibration for two classifiers and different performance metrics. The effect of calibrating a classifier is most pronounced in its *sensitivity*. There is a trade-off between specificity and sensitivity, i.e. when lowering θ_0, sensitivity increases while specificity decreases and vice versa. When θ_0 is decreased, the number of true negatives (TN) is decreased as well to a certain extent (this is undesired but unavoidable). As specificity is defined through (TN/TN+FP, see Methods) and the number of false positives stays constant, specificity decreases. *E. coli* and *S. aureus* both reach sensitivity and specificity values larger than 90% at a θ_0 value of 0.075, with rejection percentages of 43% and 44%, respectively. For *K. pneumoniae*, the depicted θ_0 and θ_1 values do not reach sensitivity and specificity both larger the 90%. However, for θ_0 of 0.025 and θ_1 of 0.8, a sensitivity of 97% and specificity of 98% can be reached at a rejection percentage of 88% (see **Suppl. Tab. 4**).

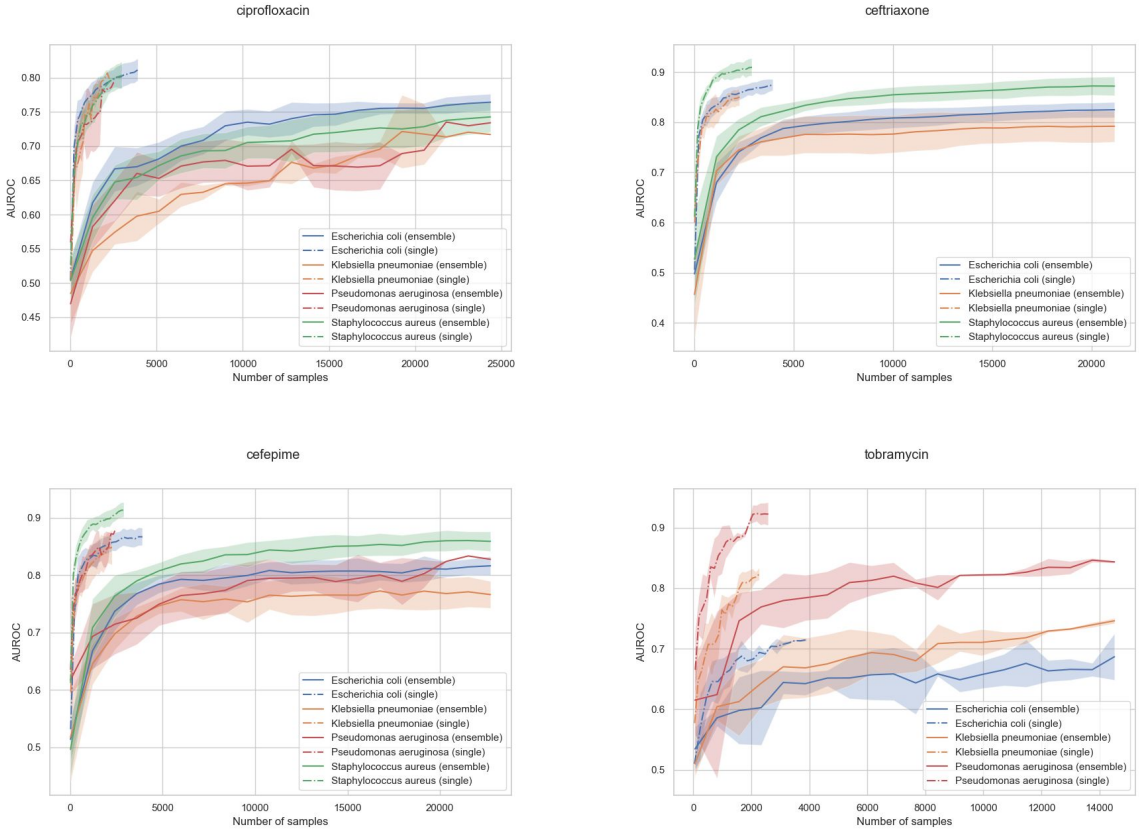


Figure 4: Predictive performance with increasing sample size AUROC with increasing sample size for complete and species-stratified *DRIAMS-A* datasets. Experiments were repeated 10 times with different shuffled train–test splits. The solid curve represents the mean of these repetitions, with the envelope depicting the standard deviation over the repetition experiments. Results are shown for several antibiotics and the four most frequent species *S. aureus, E. coli, K. pneumoniae,* and *Pseudomonas aeruginosa*. Please note that tobramycin resistance is not frequently measured in *S.aureus* and no curve is depicted (see **Suppl. Tab. 2**). The resistance profiles of *S. aureus* for ceftriaxone and cefepime are both derived through its oxacillin resistance, and are therefore the same for each single sample. These drugs are shown for comparison reasons and are usually not used to treat infections with *S. aureus*. With equal sample size, training only on samples from a single species is outperforming training on an ensemble of all available species.

Learning on a single species yields superior predictions compared to learning on multiple species

Different species of bacteria can develop resistance to a specific antimicrobial through different mechanisms. Resistance against beta-lactam antibiotics in gram-negative bacteria, such as *E. coli*, may be caused by the production of beta-lactamases such as CTX-M[24], TEM, SHV[25,26] or carbapenemases e.g. OXA-48[27]. Resistance against beta-lactam antibiotics in gram-positive bacteria, such as *S. aureus*, can be caused by a penicillinase (blaZ) resulting in a resistance only against penicillin[28] or by an alteration within the penicillin binding protein (PBP2a) resulting e.g. in the methicillin resistant *S. aureus* phenotype with resistance against multiple beta-lactamase antibiotics[29]. Therefore, pooling the spectra across species and predicting antimicrobial resistance using the same model regardless of the species poses a more complex learning task than predicting antimicrobial resistance within one specific species. However, stratification by species reduces the number of samples available for training and might lower predictive performance. We study this trade-off by comparing the performance of (i) a model trained to predict antimicrobial resistance using samples from across all species (*ensemble*) with (ii) a collection of models

trained separately for *single* bacterial species (**Fig. 4**). Each point of the depicted curves corresponds to one classifier, trained with the number of samples specified on the x-axis. The last, i.e. rightmost, of those classifiers hence corresponds to the scenario in which all available samples are being used. We observe that training a model for individual species *separately* leads to improved performance for the majority of species, despite the reduction in sample size. Notably, all training samples used to reach the last (i.e. rightmost) per-species classification results are also included in the training samples for the last ensemble classifier. The last ensemble classifier therefore has access to at least the same amount of information about the respective species as the last per-species classifier. Nevertheless, it never outperforms the per-species classifiers. Only few curves reach a plateau, with the *single* species classifier increasing more sharply with the last addition of more training samples. This indicates the higher complexity of the ensemble prediction task and the benefit of an even larger training dataset.

Site-specific and current samples are required for accurate antimicrobial resistance prediction

To guide and encourage further method development, we want to illustrate challenges and limits of mass spectra based antimicrobial resistance prediction. To this end, we study if recent samples are necessary and if adding more samples collected at older timepoints will increase prediction performance.

Fig. 5 depicts two temporal validation scenarios of training data collected at different timepoints. We first compare the predictive power of datasets collected in different years using a single-year scenario. **Fig. 5** shows a clear decrease in performance with increasing temporal distance between training and testing data. We explain this drop by two differences that accumulate over time, namely (i) biological differences through the ongoing evolution of the local microbial populations, and (ii) technical differences, such as changes after machine maintenance (e.g. laser replacement and adjustment of internal spectra processing parameters through machine calibration).

Additionally, we validate our classification approach across *multiple* sites of the *DRIAMS* datasets. First, we determine whether a model trained and tested on one *DRIAMS* site reaches similar predictive performance on other sites. Second, we study the applicability of our prediction workflow beyond the dataset collected at the *DRIAMS-A*, by training and testing the same models applied to *DRIAMS-A* to three other sites, *DRIAMS-B* to *-D*. The results for four species–antibiotic scenarios in **Suppl. Fig. 4** indicate that training on the same site as testing is beneficial for the majority, even if fewer samples are available.

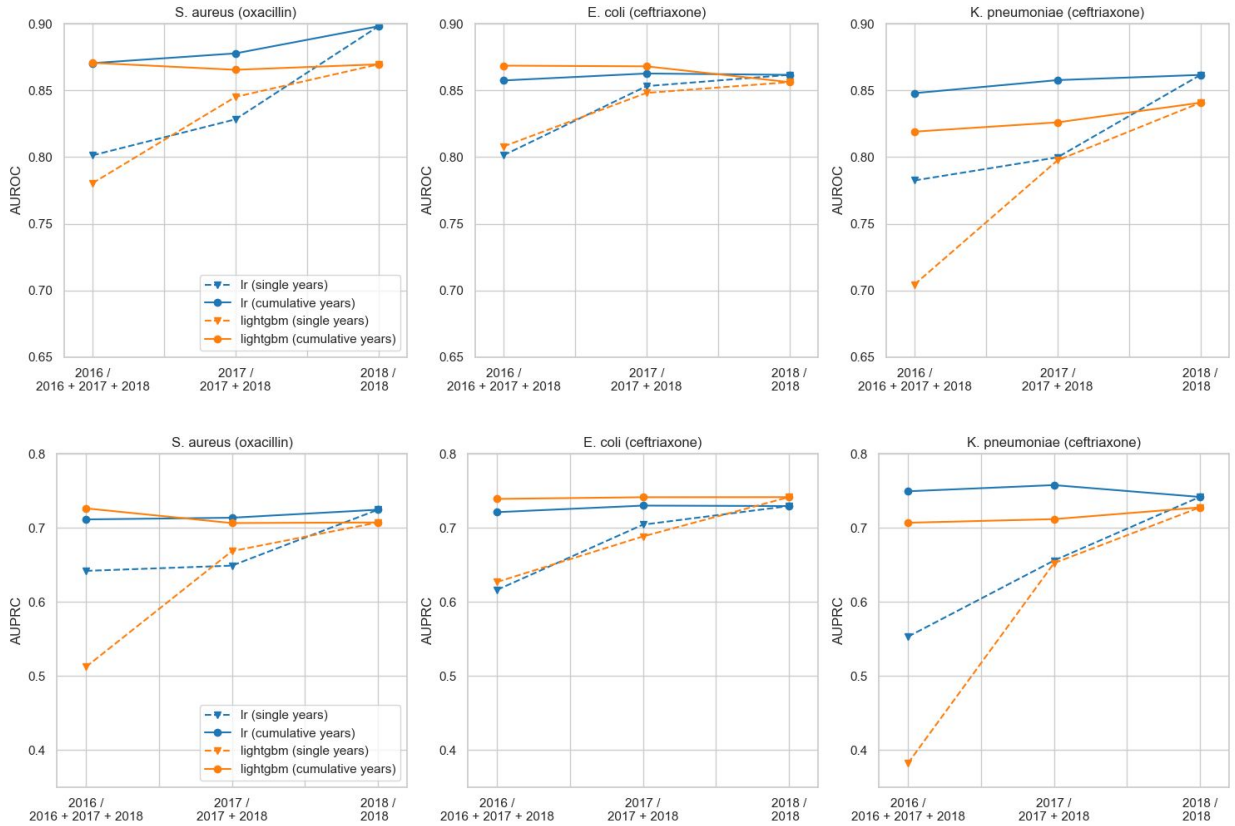


Figure 5: Temporal validation on different years of *DRIAMS-A* The figure depicts the decrease of predictive performance with increasing age of training data. As a reference, we took the most recent data from *DRIAMS-A*, collected in the year 2018. This data was split into 80% training and 20% test dataset. The test dataset was used as test data for all depicted scenarios. For the single-year training scenarios, we compare predictive performance of LR and LightGBM on three different training datasets; the 80% training data from 2018, and training datasets from 2017 and 2016, sampled to have identical sample size and class ratio as the 2018 dataset. This sampling was done to exclude different data availability or class ratios over time from affecting the predictive performance. The mean of 10 repetitions with different random splits are depicted. We observe that training data isochronous to the test data reaches the highest performance, with decreasing performance for older training data. In cumulative-year scenarios, the training datasets which were constructed for the single-year scenarios were combined. Consequently, the training data of the 2016+2017+2018 has three times the number of samples compared to 2018. The results are mixed in some cases the predictive performance benefits from a larger dataset consisting of older and isochronous samples, in other cases, predictive performance decreases.

Discussion

We demonstrate that mass spectra based antimicrobial resistance prediction from routine diagnostics is capable of providing accurate predictions with unmatched speed within 24 hours of sample collection. This analysis was made possible by collecting the largest real-world clinical dataset containing MALDI-TOF mass spectra and corresponding antimicrobial resistance profiles. Overall, we observe high performance in multiple pathogen species trained on individual species–antibiotic combinations. Using calibrated classifiers and implementing a novel rejection scenario, we were able to reach clinical utility in clinical most relevant cases such as ceftriaxone resistance of *E. coli* and *K. pneumoniae* and oxacillin resistance in *S. aureus*. We also saw that the predictive performance of classifiers trained on a single species is always higher than that of classifiers trained on ensembles. In the following, we discuss certain aspects of our observations in more detail.

We first note that for species-specific antimicrobial resistance prediction, we reach high performance with classical machine learning models such as logistic regression, obtaining AUROC values larger than 0.85 for several antibiotic-species scenarios (**Fig.3**). Generally, an antimicrobial resistance predictor guiding antimicrobial treatment decision must satisfy several requirements: (i) Excellent predictive performance, concerning sensitivity in particular. In clinical practice, assuming a bacterial infection will be susceptible (negative class) to an antibiotic while in reality, it is resistant or intermediate (positive class), has the worst impact on patient treatment efficacy: falsely assuming susceptibility results in an *ineffective* treatment. While falsely assuming resistance results in an alternative but potentially still effective treatment Thus, the *sensitivity* of the predictor should be high. (ii) The classifier must have the capability to reject those samples for which it cannot make a confident prediction. This greatly allows to adapt the predictions towards specific clinical scenarios e.g. high risk patients where a wrong treatment would have particular strong consequences. Less confident prediction is a consequence of the nature of the prediction task: We will never cover all bacterial strains occurring world-wide in the training dataset, and the strains included in training will most likely cover local endemic bacterial occurrences. When new lineages are introduced into the local bacterial landscape, through travelling for instance, a classifier cannot make an informed decision and *must* indicate so to the treating clinician. Furthermore, we have a way to proceed even if no informed decision is available, namely the treatment procedures which are currently followed. Since current clinical practice lacks additional information on antimicrobial resistance beyond the species of the infection, refusing to make a prediction cannot result in adverse effects.

As we demonstrated in **Tab. 1**, calibration followed by threshold adjustments can reach very high levels of predictive performance. The rejection percentages of classifiers required to reach high predictive performance may appear to be substantial at first glance, with rejection percentages of at least 40% required to reach specificity and sensitivity values both over 90% for scenarios *E. coli*-ceftriaxone and *S. aureus*-oxacillin. However, we argue that even a predictor with a high rejection percentage is clinically very useful for patient management, as long as the accepted samples are classified with high sensitivity and specificity, respectively. As we outlined above, our method provides guidance for antimicrobial treatment decisions and hospital epidemiological management with unmatched speed at a point in time where no information beyond the species label is available. As long as our model is accurate on the accepted samples, treatment only deviates from the current procedure if we are certain that predictions are beneficial. We do not expect adverse effects for the patients, and can provide antimicrobial treatment guidance for roughly 60% of the patients, at very low error rates.

Fig. 4 depicts that classifiers perform best for a specific scenario (e.g. a species and antibiotic), when the training dataset includes only samples from that scenario, and adding samples from other species in fact lowers the predictive performance. We distinguish two potential situations: (i) the mechanism causing antimicrobial resistance varies between species or (ii) the resistance mechanism is the same in several species. If situation (i) were true, the results in **Fig. 4** can be expected, as the classifier has to predict several objectives at once (predicting resistance caused by mechanism A in species A, while also being able to predict resistance caused by mechanism B in species B, etc.). In situation (ii), one could expect that combining samples of several species to predict resistance caused by the same mechanism increases performance.

For most scenarios, we are confronted with situation (ii), such as beta-lactam (relevant for ceftriaxone and cefepime) resistance in *E. coli* in **Fig. 4**. However, many proteins causing resistance are beyond the mass range of MALDI-TOF mass spectra, e.g. the penicillin binding protein in *S. aureus* has a mass of approximately 76'400 Da[30], beta-lactamases in *E. coli* and *K. pneumoniae* weigh approximately 30'000 Da[31–34] and the *E. coli* outer membrane porine OmpC weighs approx. 40'300 Da[35]. Previous work detecting single MALDI-TOF MS peaks associated with resistance[36,37] did not lead to peaks that could be matched to be present for the same resistance mechanism in other species. Therefore, we hypothesise that our predictor cannot detect the resistance mechanism (due to physical limitations), but rather species-specific and resistance-associated changes in the proteome as well as phylogenetic similarity between resistant vs. susceptible samples. Note that learning to predict resistance for a specific scenario (e.g. a species and antibiotic) from a training dataset combined from different species poses a higher challenge, as a much more complex set of phylogenetic relations have to be learned.

The number of samples depicted in **Fig. 4** also gives an indication at what sample size most information/variance of samples occurring at the *DRIAMS-A* collection site (University Hospital Basel) are covered by the training data. The sample sizes required to reach the 'plateau' visible for single-species predictive performance range from 2'500 to 5'000. This corresponds to roughly one to two years of data collected at the DRIAMS-A hospital, after excluding many samples as their mass spectra and resistance profile could not be matched unambiguously (see Methods). Many samples are lost through preprocessing, as clinical laboratories typically do not foresee matching of mass spectra and their corresponding laboratory analysis. We highly recommend that hospitals attempting to use mass spectra based antimicrobial resistance prediction should extend their laboratory protocols to facilitate unambiguous matching between all spectra and resistance profiles. This would ensure that the vastness of species and strains coming into clinics can be considered in the training dataset, potentially increasing both performance and confidence of the classifiers, while also keeping the percentage of rejected samples low.

We also observed that antimicrobial resistance classifiers trained on mass spectra of species from one site are not directly applicable to mass spectra measured at different sites (see **Suppl. Fig. 4**). This is influenced by many sources, including (i) different phylogenetic strains, (ii) different prevalences of resistance (i.e. different class ratios), which can impact predictive performance, or (iii) technical variability[38], due to different machine-specific parameters and settings. In a similar vein, the closer the time of collecting the training samples is to the time of prediction, the better the predictive power of the trained classifier (see **Fig. 5**). Again, possible sources include accumulating differences in phylogenetic strains and continuous changes in technical conditions. Critically, this implies that a clinically-applied classifier must be retrained with the most recent data in regular time intervals.

Many exciting unexplored avenues remain in MALDI-TOF based antimicrobial resistance prediction. Future experiments could explore the biological samples based on their feature importance values, obtained either directly during training (for a logistic regression classifier) or calculated via interpretable prediction scores, such as Shapley values[39]. Feature importance values help gauge the relevance of MALDI-TOF MS peaks that are assigned to a specific bin. Through techniques such as MALDI-TOF/TOF MS, proteins or protein fragments generating a MALDI-TOF MS peak can be captured, sequenced and identified. By analysing the content of relevant feature bins, a broader understanding of the information used for the high predictive performance can be achieved. Analysing the reason for some samples receiving a low probability score could be beneficial for increasing the number of high-confidence samples. While we expect samples of strains underrepresented in the training dataset to receive a low confidence, analysing the influence of spectra quality on the confidence could be beneficial to increasing the number of non-rejected samples. Another open question is whether advanced machine learning models for transfer learning scenarios can overcome the cross-site validation challenges reported here and either reduce or remove the need for retraining; we plan to explore this question in future work.
In summary, our work demonstrates that MALDI-TOF MS based machine learning is a research field of the utmost importance and can provide radical new ways to predict antimicrobial resistance in clinical most relevant scenarios, with idiosyncratic challenges that need to be overcome, but also a large potential to improve patient treatment.

Online Methods

Reproducibility of Results and Data availability

All R and Python scripts can be found in https://github.com/BorgwardtLab/maldi_amr. The full dataset can be downloaded at the open science foundation https://osf.io/fjr4s/?view_only=d21decba094445a2ae46704e57ed9db2.

MALDI-TOF mass spectra acquisition and antimicrobial resistance testing

We collected data from daily clinical routine at ISO/IEC 17025 accredited diagnostic routine laboratories. The study was evaluated by the local ethical committee (IEC 2019-00729). Specifically, all MALDI-TOF mass spectra contained in *DRIAMS-A* to *-F* were acquired at four microbiological laboratories in Switzerland providing routine diagnostic services for hospitals and private practises. All laboratories use the Microflex Biotyper System by Bruker Daltonics (Bremen, Germany), which is a widely-employed MALDI-TOF MS systems in microbiological routine diagnostics both in North America[16] and in Europe[17,18]. The four diagnostic laboratories included in this study are (1) University Hospital Basel-Stadt (providing *DRIAMS-A*), (2) Canton Hospital Basel-Land (providing *DRIAMS-B*), (3) Canton Hospital Aarau (providing *DRIAMS-C*), and (4) laboratory service provider Viollier (providing *DRIAMS-D*). While Canton Hospitals Basel-Land and Aarau employ the Microflex Biotyper LT/SH System, Viollier uses the Microflex smart LS System. Although these two systems differ in their respective laser gas, they include the same reference spectra database, so we included spectra of both Microflex Biotyper systems. University Hospital Basel-Stadt uses both Microflex Biotyper systems in parallel. The species of each mass spectrum was identified using the Microflex Biotyper Database (MBT 7854 MSP Library, BDAL V8.0.0.0_7311-7854 (RUO)) included in the flexControl Software (Bruker Daltonics flexControl v.3.4). Similar to the mass spectra, antimicrobial resistance profiles were routinely acquired in the same four microbiological laboratories within the same time frames of the dataset. Resistance categories for bacteria values were determined either using microdilution assays (VITEK® 2, BioMérieux, Marcy-l'Étoile, France), or by minimal inhibitory concentration (MIC) stripe tests (Liofilchem, Roseto degli Abruzzi, Italy), or disc diffusion tests (ThermoFisher Scientific, Waltham, USA). Resistance categories for yeast were determined by using Sensititre Yeast One (Thermofisher). All breakpoint measurements were interpreted to be either susceptible, intermediate, or resistant according to EUCAST [40] and CLSI (2015 M45; 2017 M60) recommendations. The EUCAST versions used were updated with every EUCAST Breakpoints table update and include v6-v8.

Quality control

Empty spectra and calibration spectra were excluded from further analysis. This serves to ensure a similar level of data quality for the different sites.

Matching of MALDI-TOF mass spectra and antimicrobial resistance profiles

MALDI-TOF MS based antimicrobial resistance prediction requires a dataset containing mass spectra and their corresponding resistance labels, in the form of antimicrobial resistance profiles. In order to construct such a dataset, MALDI-TOF MS and resistance profile measurements belonging to the same microbial isolate have to be matched. Since each site in *DRIAMS* stores the mass spectra and their corresponding antimicrobial resistance profiles in separate databases,a matching procedure has to be developed for each site.

We use the term 'laboratory report' for the document used to report laboratory measurement results, including antimicrobial resistance profiles, for each patient within the clinical care. The species of the specimen is obtained through Bruker Microflex MALDI-TOF MS and added to the laboratory report. This decouples laboratory report entry and the mass spectrum; there is no link required between the spectrum file and the laboratory entry after the species is entered. The antimicrobial resistance profiles obtained in their individual experiments are also added to the laboratory report. The laboratory report entries are commonly identified by codes linking them to a patient, or a unique sample taken from a patient, to which we refer as 'sample ID'. Multiple entries with the same sample ID can exist, if several probes were taken from the same patient or several colonies tested from the same probe.
In general, the spectra recorded by the Bruker Microflex systems were labelled with an ambiguous, i.e. non-unique, code corresponding to the non-unique sample ID in the laboratory report. MALDI-TOF mass spectra and their

corresponding antimicrobial resistance profiles were stored in separate files. In the clinic, MALDI-TOF MS spectra are never intended to be matched up with the laboratory report entries; therefore no proper protocols for matching exist. Matching protocols had to be developed uniquely and in an ad-hoc fashion for each labelling system at each institution.

In order to link mass spectra to their antimicrobial resistance profiles, we constructed a *unique* identifier, using the sample ID and the determined genus of a sample. The rationale behind this strategy is that if multiple sample ID entries exist, this is most likely due to multiple genera being present in the patient samples, leading to several measurements. We omitted samples for which we were unable to construct a unique sample ID–genus pair.
Mass spectra were stored without information on the determined species. Hence, for each spectrum, the species and genus label is determined by re-analysing the spectra with the University Hospital Basel-Stadt Bruker library and then matching the spectrum to its corresponding antimicrobial resistance profile using the assigned sample ID and the determined genus. All MALDI-TOF MS systems used in this study were maintained according to the manufacturer's standard and spectra were routinely acquired using the 'AutoXecute' acquisition mode. The genus is used (instead of species), as it allows for some flexibility between the species assigned to a sample in the laboratory report and the Microflex Biotyper. The species label given in the laboratory report can differ from the species assigned to the corresponding MALDI-TOF mass spectrum by the Microflex Biotyper System as additional microbiological tests can give a more accurate label. In what follows, we provide additional details regarding the matching procedure which are specific to each site.

University Hospital Basel-Stadt Starting in 2015, the spectra were labelled with a 36-position code by the Bruker machine (e.g. 022b130c-6c8c-49b5-814d-c1ea8b2e7f93), which we term 'Bruker ID'. This code is guaranteed to be unique for all spectra labelled from one machine. Each AMR profile is labelled with a 6-digit sample ID, which is unique for samples in one year. AMR profiles were collected using the laboratory information system. The laboratory information system includes all entries made for a sample, also false entries which have later been corrected and have not been reported nor considered for patient treatment. As such manual corrections are very rare, the uncertainty in AMR labelles is however limited. For each year (2015, 2016, 2017, and 2018) there are separate antimicrobial resistance profile tables and folders containing all spectrum samples collected during the corresponding year. We lost 40'569 spectra out of 186'098 by following the aforementioned pre-processing routines (*DRIAMS-A*).

Canton Hospital Basel-Land The antimicrobial resistance profiles and mass spectra are each labelled with a 6-digit sample ID. The genus depicted in each mass spectrum was determined through comparison to the Microflex Biotyper Database (Bruker Daltonics flexControl v.3.4); the genus of each antimicrobial resistance profile was stated in the laboratory report. Mass spectra and AMR profiles were merged using the 6-digit sample ID and the genus information.

Canton Hospital Aarau Here, the laboratory report contains the 10-digit sample ID, species label, and antimicrobial resistance profiles of measured samples. This software version did not provide a unique 36-character code for each spectrum, but only a 10-digit sample ID that had to be used to match spectra to the antimicrobial resistance profiles from the laboratory. Since the sample ID can be shared by different spectra, it is insufficient to uniquely match a species label to an input spectrum. To circumvent this problem, we divided the spectra in 15 batches, each one only containing unique 10-digit sample IDs. Repeated sample IDs were distributed over the batches. These 15 batches were re-analysed and labelled by the Bruker software, and 15 output files with the given species labelled were created. Through the separation in batches, the certain species label was determined for each spectrum. The label for each spectrum in the batches can be determined, as we only included spectra that already had a label in the lab file. Now, each spectrum file has a combined label made up of its 10-digit sample ID and its species label. If this combined label was found to have a unique match within the lab results file, the AMR profile was assigned to the spectrum, otherwise its AMR profile position remains empty and only the spectrum with its species label was added to the dataset. We ignore all spectra that could not be matched to an entry in the lab results file (such spectra arise from measurements that do not provide AMR information).

Viollier While all other sites reported AMR labels with either 'R', 'S', 'I', 'positive' or 'negative' values, samples provided by Viollier are labelled with precursory measurements, namely the MIC of each antibiotic. We therefore use the breakpoints given the up-to-date EUCAST guidelines (v.9) to convert the MIC values to 'RSI' values.
80,'796 spectra in the `fid` file format are present, identified through a unique 36-character 'Bruker ID'. The AMR results are identified by a 10-digit sample ID, which are linked to the Bruker IDs in an additional file, the 'linking file'. The main reasons for loss of data in preprocessing are (1) the AMR results and ID 'linking files' contained significantly fewer entries than `fid` files present (40'571 and 51'177 respectively) and (2) following advice by the lab personnel, only the 10-digit sample ID could be used for matching to the BrukerID (which contained a longer version of the LabID). Through exclusion of all entries without a unique 10-digit sample ID in both the AMR results and linking files, another significant portion of data was lost. Specifically, there is an overlap of 10,852 filtered entries from the laboratory report file and the linking file. After matching these entries with spectra, 7,771 spectra with 7,720 AMR profiles remained. Spectra without an AMR profile are not used for any supervised learning tasks (such as prediction).

Conversion to *DRIAMS*

We require a DRIAMS dataset entry to contain (1) a spectra code linking unambiguously to a spectra file, (2) a readable mass spectra `fid` file (as all of our mass spectra are measured on Bruker Microflex machines), (3) the corresponding species (species reported as 'Organism best match' by the flexControl Software was assigned), and, if available, (4) AMR profiles stating antimicrobial susceptibility results.

Antimicrobial nomenclature

Since the naming scheme of antimicrobial drugs was not consistent between sites, due to different spelling variants or the use of several names for the same drug (e.g. cotrimoxazol and trimethoprim/sulfamethoxazole describe the same drug), we unified them during preprocessing. Specifically, we anglicized the names of antimicrobial drugs from their original German version in the DRIAMS ID files as a preprocessing step for our machine learning analysis.
Antimicrobial names were specified by adding suffixes to their names. In the following, we explain the used suffix nomenclature for DRIAMS-A: (i) 'high level': According to EUCAST, MIC breakpoints vary with the dosage of Gentamycin (standard: 5mg / kg and high level: 7 mg / kg for i.v. administration); (ii) 'meningitis', 'pneumoniae', 'endocarditis', 'uncomplicated urinary tracts infections (UTI)': EUCAST includes infection specific breakpoints for these infections (see suffixes for amoxicillin-clavulanic acid, penicillin and meropenem); (iii) 'screen': cefoxitin is used to screen for MRSA in clinical routine diagnostic at University Hospital Basel, using the respectively-defined breakpoints by EUCAST; (iv) 'GRD': GRD (glycopeptide resistance detection) is a MIC Strip Test used at University Hospital Basel in very rare cases to detect glycopeptide intermediate *S. aureus*; (v) '1mg_l' indicates the concentration of rifampicin, when MIC are measured in liquid culture as it is routinely done for *Mycobacterium tuberculosis* (MTB). These entries of non-MTB species were entered incorrectly into the laboratory information system.

Dataset characteristics

All medical institutions are located in Switzerland. Microbial samples in the University Hospital Basel-Stadt database (i.e. *DRIAMS-A*) mostly originate from patients located in the city of Basel and its surroundings. Such patients visit the hospital for either out- or inpatient treatment. Samples in the Canton Hospital Basel-Land dataset (i.e. *DRIAMS-B*) primarily originate from the town surrounding the City of Basel. Patients from the Swiss Canton Aargau seek medical care at the Canton Hospital Aarau (*DRIAMS-C*). Viollier (*DRIAMS-D*) is a service provider that performs species identification for microbial samples collected in medical practices and hospitals. Samples originate from private practices and hospitals all over Switzerland.

DRIAMS-A to *-D* are datasets that contain data collected in the daily clinical routine. All mass spectra measured in a certain time frame are included. The time frame during which each dataset was collected are as follows:

DRIAMS-A: 34 months (11/2015–08/2018)
DRIAMS-B: 6 months (01/2018–06/2018)
DRIAMS-C: 8 months (01/2018–08/2018)
DRIAMS-D: 6 months (01/2018–06/2018)

Spectral representation

In the *DRIAMS* dataset, we include mass spectra in their raw version without any preprocessing, and binned with several bin sizes. For all machine learning analysis in this study, a bin size of 3 Da was used. The spectra are extracted from the Bruker Flex machine in the Bruker Flex data format. The following preprocessing steps were performed using the R package `MaldiQuant`[41] version 1.19: (1) the measured intensity was transformed with a square-root method to stabilize the variance, (2) smoothing using the Savitzky-Golay algorithm with half-window-size 10 was applied, (3) an estimate of the baseline was removed in 20 iterations of the SNIP algorithm, (4) the intensity was calibrated using the total-ion-current (TIC), and (5) the spectra were trimmed to values in a 2’000 to 20’000 Da range. For exact parameter values please refer to the code.

After preprocessing, each spectrum is represented by a set of measurements, each of them described by its corresponding mass-to-charge ratio and intensity. However, this representation results in each sample having potentially a different dimensionality (i.e. cardinality) and different measurements being generally irregularly-spaced. Since the machine learning methods used in this manuscript require their input to be a feature vector of fixed dimensionality, intensity measurements were binned using a bin size of 3 Da. To perform the binning, we partition the m/z axis in the range of 2’000 to 20’000 Da into disjoint, equal-sized bins and sum the intensity of all measurements in the sample (i.e. a spectrum) falling into the same bin. Thus, each sample is represented by a vector of fixed dimensionality, i.e. a vector containing 6’000 features, which is the number of bins the m/z axis was partitioned into. We use this feature vector representation for all downstream machine learning tasks.

Antimicrobial resistance phenotype binarization

For the machine learning analysis, the values of antimicrobial resistance profiles were binarized during data input to have a binary classification scenario. The categories are based on EUCAST and CLSI recommendations. For tests that report RSI values, resistant (R) and intermediate (I) samples were labelled as class 1, while susceptible (S) samples were labelled as class 0. We grouped samples in the intermediate class together with resistant samples, as both types of samples prevent the application of the antibiotic. In the EUCAST v6-v8 the intermediate category shows higher MIC values but due to safety reasons in clinical practice this was usually counted to resistance in order to have safty buffer in reaching sufficient high antibiotic drug concentrations.

Machine learning methods

For AMR classification, we used numerous state-of-the-art classification algorithms with different capabilities. This included (1) logistic regression, (2) random forests[42], (3) support vector machines[43] with different kernels, and (4) LightGBM[44], a modern variant of gradient-boosted decision trees. For LightGBM, we use the official implementation in the `lightgbm` package, while we use the scikit-learn package for all other models[45].

For each antibiotic, all samples with a missing AMR profile of the antibiotic were removed and the machine learning pipeline was applied to the reduced dataset. Samples were randomly split into a training dataset comprising 80% of the samples and a test dataset with the remaining 20%, while stratifying for *both* the class and the species of the samples. This is slightly unusual in standard machine learning setups, which typically only require stratification by a single class label, but crucial for our scenarios to guarantee similar prevalence values. To select an appropriate model configuration for a specific task, we employ 5-fold cross-validation; in case an insufficient number of samples is available, our implementation falls back to a 3-fold cross-validation on the training dataset to optimise the respective hyperparameters. Hyperparameters are model-specific (see below for more details), but always include the choice of an optional standardisation step (in which feature vectors are transformed to have zero mean and unit variance). To determine the best-performing hyperparameter set, we optimised the area under the ROC curve (AUROC). This metric is advantageous in our scenario, as it is not influenced by the class ratio and summarises the performance of correct and incorrect susceptibility predictions over varying classification score thresholds. Having selected the best hyperparameters, we retrain each model on the full training dataset, and use the resulting classifier for all subsequent predictions. Our hyperparameter grid is extensive, comprising, for example, the choice of different logistic regression penalties (L1, L2, no penalty), the choice of scaling method (standardisation or none), and regularisation parameters ($C \in \{10^{-3}, 10^{-2}, \cdots 10^{2}, 10^{3}\}$). For more details, please refer to our code (`models.py`).
We implement all models in Python and publish them in a single package, which we modelled after scikit-learn, a powerful library for machine learning in Python.

Evaluation metrics

We report AUROC as the main metric of performance evaluation. The datasets of most antibiotics under consideration exhibit a high class imbalance (20 out of 42 antibiotics show a resistant/intermediate class ratio <20% or >80%). AUROC is invariant to the class ratio of the dataset and therefore permits a certain level of comparability between antibiotics with different class ratios. A pitfall of reporting AUROC in the case of unbalanced datasets, however, is that it does not reflect the performance with respect to precision (or positive predictive value). Therefore, the AUROC value can be large while precision is low, and it is tempting to be overly optimistic about the AUROC values. To account for this bias, we additionally report the area under the precision-recall curve (AUPRC) values; this metric is not used during the training process, though.
Two other metrics commonly used in clinical research are sensitivity and specificity. Sensitivity measures to what extent positives are recognized and specificity measures how well negatives are detected as such. Analogous to the AUROC curves, we show sensitivity vs. specificity curves to illustrate the tradeoff between both metrics. Please note commonalities to other metrics; *sensitivity*, *recall* and *true positive rate* are synonyms and all correspond to the same metric; specificity is a counterpart to the false positive rate; $\textit{true positive rate} = 1 - \textit{specificity}$.

Connection to confusion matrix All of the metrics we employ here can be derived from the counts within a confusion matrix. The number of true positives (TP) is the number of test data points that are classified as resistant or intermediate by the classifier and confirmed resistant or intermediate in a phenotypic antimicrobial susceptibility test. True negatives (TN) is the number of test data points correctly classified as belonging to the susceptibility class. The false positive count (FP) refers to the number of susceptible test data points classified to be resistant/intermediate, whereas false negatives (FN) are the number of resistant/intermediate test data points classified to be susceptible.
The area under the receiver operating characteristic (AUROC) shows the true positive rate (TP/TP+FN) against the false positive rate (FP/FP+TN). The AUPRC, as well as the AUROC, is traditionally reported on the minority class. In our scenario, however, while the minority class is the resistant class in most cases, this is not consistent and for some antibiotics more samples of the resistant will be present. The precision-recall curve shows the recall

(TP/TP+FN) against the precision (TP/TP+FP). A perfect classifier would show a performance of 1.0 for each AUROC and AUPRC. The performance of a random classifier would be 0.5 for AUROC and percentage of samples of the positive (susceptible) class for AUPRC. As mentioned above, sensitivity and susceptibility are synonyms for recall and precision, respectively.

Acknowledgements We thank Olivia Grüninger, Josiane Reist, Yoann Mahiddine, Daniela Lang, Clarisse Straub and Magdalena Schneider for excellent technical assistance in accessing the mass spectra and antimicrobial resistance data at the University Hospital Basel. We further thank Belén Rodríguez-Sánchez (Hospital General Universitario Gregorio Marañón, Madrid, Spain), Helena Seth-Smith (USB), Carole Kaufmann (USB) and Valentin Pflüger (MabritecAG, Riehen, Switzerland) and Maximilian Brackmann (Federal Office for Civil Protection, Spiez Laboratory, Bacteriology, Spiez, Switzerland) for valuable advice, technical consulting, and feedback throughout the project.

Competing Interests statement
The authors declare no competing interests.

Funding This study was supported by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung (Borgwardt, Rieck), ETH personalised medicine grant (PMB-03-17, Weis)

References:

1. Global action plan on antimicrobial resistance. https://www.who.int/publications/i/item/global-action-plan-on-antimicrobial-resistance.
2. Wise, R. *et al.* Antimicrobial resistance. Is a major threat to public health. *BMJ* **317**, 609–610 (1998).
3. Cassini, A. *et al.* Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **19**, 56–66 (2019).
4. Kumar, A. *et al.* Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* **34**, 1589–1596 (2006).
5. Seymour, C. W. *et al.* Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. *N. Engl. J. Med.* **376**, 2235–2244 (2017).
6. Huang, A. M. *et al.* Impact of rapid organism identification via matrix-assisted laser desorption/ionization time-of-flight combined with antimicrobial stewardship team intervention in adult patients with bacteremia and candidemia. *Clin. Infect. Dis.* **57**, 1237–1245 (2013).
7. Osthoff, M. *et al.* Impact of MALDI-TOF-MS-based identification directly from positive blood cultures on patient management: a controlled clinical trial. *Clin. Microbiol. Infect.* **23**, 78–85 (2017).
8. Banerjee, R. *et al.* Randomized Trial of Rapid Multiplex Polymerase Chain Reaction-Based Blood Culture Identification and Susceptibility Testing. *Clin. Infect. Dis.* **61**, 1071–1080 (2015).
9. Kommedal, Ø., Aasen, J. L. & Lindemann, P. C. Genetic antimicrobial susceptibility testing in Gram-negative sepsis--impact on time to results in a routine laboratory. *APMIS* **124**, 603–610 (2016).
10. Core Elements of Antibiotic Stewardship | Antibiotic Use | CDC. https://www.cdc.gov/antibiotic-use/core-elements/index.html (2019).
11. van Belkum, A., Welker, M., Pincus, D., Charrier, J. P. & Girard, V. Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry in Clinical Microbiology: What Are the Current Issues? *Ann. Lab. Med.* **37**, 475–483 (2017).
12. Croxatto, A., Prod'hom, G. & Greub, G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews* vol. 36 380–407 (2012).
13. Dierig, A., Frei, R. & Egli, A. The fast route to microbe identification: matrix assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS). *Pediatr. Infect. Dis. J.* **34**, 97–99 (2015).

14. Gradel, K. O. *et al.* Impact of appropriate empirical antibiotic treatment on recurrence and mortality in patients with bacteraemia: a population-based cohort study. *BMC Infect. Dis.* **17**, 122 (2017).
15. Retamar, P. *et al.* Impact of inadequate empirical therapy on the mortality of patients with bloodstream infections: a propensity score-based analysis. *Antimicrob. Agents Chemother.* **56**, 472–478 (2012).
16. Kim, J.-M. *et al.* Rapid Discrimination of Methicillin-Resistant Staphylococcus aureus by MALDI-TOF MS. *Pathogens* **8**, (2019).
17. Weis, C. *et al.* Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. *Bioinformatics* **36**, i30–i38 (2020).
18. Vervier, K., Mahé, P., Veyrieras, J.-B. & Vert, J.-P. Benchmark of structured machine learning methods for microbial identification from mass-spectrometry data. *arXiv [stat.ML]* (2015).
19. Weis, C. V., Jutzeler, C. R. & Borgwardt, K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. *Clin. Microbiol. Infect.* (2020) doi:10.1016/j.cmi.2020.03.014.
20. Wang, H.-Y. *et al.* A large-scale investigation and identification of methicillin-resistant Staphylococcus aureus based on peaks binning of matrix-assisted laser desorption ionization-time of flight MS spectra. *Briefings in Bioinformatics* (2020) doi:10.1093/bib/bbaa138.
21. World Health Organization (WHO). WHO publishes list of bacteria for which new antibiotics are urgently needed. https://www.who.int/news-room/detail/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed (2017).
22. Centers for Disease Control and Prevention (U.S.). Antibiotic resistance threats in the United States, 2019. (2019) doi:10.15620/cdc:82532.
23. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. in *Proceedings of the 22nd International Conference on Machine Learning* 625–632 (Association for Computing Machinery, 2005).
24. Bevan, E. R., Jones, A. M. & Hawkey, P. M. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *J. Antimicrob. Chemother.* **72**, 2145–2155 (2017).
25. Pietsch, M. *et al.* Molecular characterisation of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli isolates from hospital and ambulatory patients in Germany. *Vet. Microbiol.* **200**, 130–137 (2017).
26. Kim, Y.-K. *et al.* Bloodstream infections by extended-spectrum beta-lactamase-producing Escherichia coli and Klebsiella pneumoniae in children: epidemiology and clinical outcome. *Antimicrob. Agents Chemother.* **46**,

1481–1491 (2002).

27. Potron, A., Poirel, L., Rondinaud, E. & Nordmann, P. Intercontinental spread of OXA-48 beta-lactamase-producing Enterobacteriaceae over a 11-year period, 2001 to 2011. *Euro Surveill.* **18**, (2013).
28. Pereira, L. A., Harnett, G. B., Hodge, M. M., Cattell, J. A. & Speers, D. J. Real-time PCR assay for detection of blaZ genes in Staphylococcus aureus clinical isolates. *J. Clin. Microbiol.* **52**, 1259–1261 (2014).
29. Long, S. W. *et al.* PBP2a mutations causing high-level Ceftaroline resistance in clinical methicillin-resistant Staphylococcus aureus isolates. *Antimicrob. Agents Chemother.* **58**, 6668–6674 (2014).
30. Beta-lactam-inducible penicillin-binding protein. https://www.uniprot.org/uniprot/P07944.
31. Beta-lactamase OXA-1 precursor. https://www.uniprot.org/uniprot/P13661.
32. Beta-lactamase TEM precursor. https://www.uniprot.org/uniprot/P62593.
33. Beta-lactamase SHV-24 precursor. https://www.uniprot.org/uniprot/Q9S169.
34. Beta-lactamase CTX-M-1 precursor. https://www.uniprot.org/uniprot/P28585.
35. ompC - Outer membrane porin C precursor. https://www.uniprot.org/uniprot/P06996.
36. Lau, A. F. *et al.* A Rapid Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry-Based Method for Single-Plasmid Tracking in an Outbreak of Carbapenem-Resistant Enterobacteriaceae. *Journal of Clinical Microbiology* vol. 52 2804–2812 (2014).
37. Rhoads, D. D., Wang, H., Karichu, J. & Richter, S. S. The presence of a single MALDI-TOF mass spectral peak predicts methicillin resistance in staphylococci. *Diagn. Microbiol. Infect. Dis.* **86**, 257–261 (2016).
38. Oberle, M. *et al.* The Technical and Biological Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Based Typing: Employment of Bioinformatics in a Multicenter Study. *PLoS One* **11**, e0164260 (2016).
39. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
40. ESCMID-European Society of Clinical Microbiology & Diseases, I. EUCAST: Clinical breakpoints and dosing of antibiotics. http://www.eucast.org/clinical_breakpoints/.
41. Jacoby, G. A. & Sutton, L. beta-Lactamases and beta-lactam resistance in Escherichia coli. *Antimicrob. Agents Chemother.* **28**, 703–705 (1985).
42. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
43. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* vol. 20 273–297 (1995).

44. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).

45. Pedregosa, F., Varoquaux, G. & Gramfort, A. Scikit-learn: Machine learning in Python. *the Journal of machine* (2011).

SUPPLEMENT

antimicrobial class	number of samples in class	average positive sample ratio	antimicrobials
OTHER BETA-LACTAM ANTIBACTERIALS	174826	0.23	cefuroxime, meropenem, ertapenem, aztreonam, ceftazidime, cefazolin, cefpodoxime, cefoxitin_screen, cefepime, meropenem_with_pneumonia, meropenem_with_meningitis, meropenem_without_meningitis, ceftriaxone, imipenem, cefixime
BETA-LACTAM ANTIBACTERIALS, PENICILLINS	107128	0.5	oxacillin, ampicillin-amoxicillin, penicillin_with_meningitis, amoxicillin-clavulanic acid_uncomplicated_hwi, piperacillin-tazobactam, penicillin, penicillin_without_endokarditis, penicillin_with_pneumonia, penicillin_with_other_infections, penicillin_with_endokarditis, amoxicillin, amoxicillin-clavulanic acid
OTHER ANTIBACTERIALS	83998	0.102	linezolid, colistin, daptomycin, vancomycin, fosfomycin-trometamol, metronidazole, fosfomycin, vancomycin_grd*, teicoplanin, fusidic acid, teicoplanin_grd*, nitrofurantoin
QUINOLONE ANTIBACTERIALS	58054	0.214	quinolones, ciprofloxacin, norfloxacin, moxifloxacin, levofloxacin
AMINOGLYCOSIDE ANTIBACTERIALS	46799	0.109	gentamicin, gentamicin_high_level, aminoglycosides, amikacin, tobramycin
SULFONAMIDES AND TRIMETHOPRIM	26640	0.183	cotrimoxazole
MACROLIDES, LINCOSAMIDES AND STREPTOGRAMINS	23099	0.357	azithromycin, erythromycin, clarithromycin, clindamycin
TETRACYCLINES	22377	0.142	tetracycline, minocycline, tigecycline, doxycycline
DRUGS FOR TREATMENT OF TUBERCULOSIS	10976	0.049	rifampicin_1mg-l**, rifampicin
ANTIMYCOTICS FOR SYSTEMIC USE	5594	0.17	5-fluorocytosine, caspofungin, itraconazole, amphotericin b, posaconazole, voriconazole, anidulafungin, fluconazole, micafungin
ANTIBIOTICS FOR TOPICAL USE	3815	0.007	mupirocin
AMPHENICOLS	203	0.138	chloramphenicol
total	**563509**	**0.244**	

Supplemental Table 1: Categorisation into antimicrobial class, average number of spectra and average resistance positive class ratio of 71 antimicrobials contained in *DRIAMS-A*. The values for each individual antimicrobial can be found in **Suppl. Tab. 2**. * vancomycin_grd and teicoplanin_grd is a MIC Strip Test used at University Hospital Basel in very rare cases to detect glycopeptide intermediate *S. aureus*; ** '1mg_l' indicates the concentration of rifampicin, when MIC are measured in liquid culture as it is routinely done for *Mycobacterium tuberculosis* (MTB). These entries of non-MTB species were entered incorrectly into the laboratory information system.

antibiotic	number of samples	positive sample ratio	most frequent species
ciprofloxacin	30543	0.244	*Escherichia coli* (4911), *Staphylococcus epidermidis* (4679), *Staphylococcus aureus* (3757), *Pseudomonas aeruginosa* (3153), *Klebsiella pneumoniae* (2838)
meropenem	29531	0.174	*Escherichia coli* (4928), *Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3643), *Pseudomonas aeruginosa* (3183), *Klebsiella pneumoniae* (2855)
imipenem	29391	0.234	*Escherichia coli* (4923), *Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3640), *Klebsiella pneumoniae* (2847), *Pseudomonas aeruginosa* (2377)
cefepime	28476	0.229	*Escherichia coli* (4890), *Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3640), *Pseudomonas aeruginosa* (3088), *Klebsiella pneumoniae* (2839)
piperacillin-tazobactam	28398	0.231	*Escherichia coli* (4799), *Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3640), *Pseudomonas aeruginosa* (3152), *Klebsiella pneumoniae* (2759)
ampicillin-amoxicillin	26871	0.817	*Escherichia coli* (4866), *Staphylococcus epidermidis** (4371), *Staphylococcus aureus** (3556), *Klebsiella pneumoniae* (2856), *Enterobacter cloacae* (1257)
cotrimoxazole	26640	0.183	*Escherichia coli* (4888), *Staphylococcus epidermidis* (4545), *Staphylococcus aureus* (3741), *Klebsiella pneumoniae* (2854), *Enterobacter cloacae* (1257)
ceftriaxone	26545	0.275	*Escherichia coli* (4961), *Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3640), *Klebsiella pneumoniae* (2860), *Enterobacter cloacae* (1249)
amoxicillin-clavulanic acid	25228	0.393	*Escherichia coli* (4826), *Staphylococcus epidermidis** (4262), *Staphylococcus aureus** (3640), *Klebsiella pneumoniae* (2840), *Enterobacter cloacae* (1260)
levofloxacin	20784	0.191	*Escherichia coli*** (4858), *Klebsiella pneumoniae*** (2830), *Pseudomonas aeruginosa*** (2356), *Enterobacter cloacae*** (1256), *Enterococcus faecium*** (1133)
colistin	18333	0.155	*Escherichia coli* (4930), *Pseudomonas aeruginosa* (3234), *Klebsiella pneumoniae* (2854), *Enterobacter cloacae* (1257), *Proteus mirabilis**** (912)
tobramycin	18190	0.093	*Escherichia coli* (4876), *Pseudomonas aeruginosa* (3231), *Klebsiella pneumoniae* (2846), *Enterobacter cloacae* (1257), *Proteus mirabilis* (912)
ceftazidime	17392	0.141	*Escherichia coli* (4822), *Klebsiella pneumoniae* (2832), *Pseudomonas aeruginosa* (2459), *Enterobacter cloacae* (1234), *Proteus mirabilis* (909)
amikacin	17222	0.057	*Escherichia coli* (4858), *Klebsiella pneumoniae* (2830), *Pseudomonas aeruginosa* (2372), *Enterobacter cloacae* (1257), *Proteus mirabilis* (903)
vancomycin	15076	0.012	*Staphylococcus epidermidis* (4777), *Staphylococcus aureus* (3791), *Enterococcus faecium* (1183), *Enterococcus faecalis* (785), *Staphylococcus hominis* (717)
ertapenem	14753	0.02	*Escherichia coli* (4983), *Klebsiella pneumoniae* (2859), *Enterobacter cloacae* (1249), *Proteus mirabilis* (915), *Serratia marcescens* (860)
penicillin	13406	0.737	*Staphylococcus epidermidis* (4371), *Staphylococcus aureus* (3553), *Staphylococcus hominis* (709), *Haemophilus influenzae* (360), *Staphylococcus capitis* (318)
linezolid	12288	0.001	*Staphylococcus epidermidis* (4407), *Staphylococcus aureus* (3639), *Enterococcus faecium* (1124), *Enterococcus faecalis* (763), *Staphylococcus hominis* (709)
tigecycline	12280	0.004	*Staphylococcus epidermidis* (4365), *Staphylococcus aureus* (3640), *Enterococcus faecium* (1128), *Enterococcus faecalis* (765), *Staphylococcus hominis* (700)
clindamycin	11612	0.313	*Staphylococcus epidermidis* (4192), *Staphylococcus aureus* (3575), *Staphylococcus hominis* (685), *Staphylococcus capitis* (318), *Staphylococcus lugdunensis* (305)
daptomycin	11384	0.009	*Staphylococcus epidermidis* (4761), *Staphylococcus aureus* (3780), *Staphylococcus hominis* (715), *Staphylococcus capitis* (365), *Enterococcus faecium* (357)
erythromycin	11079	0.409	*Staphylococcus epidermidis* (4232), *Staphylococcus aureus* (3598), *Staphylococcus hominis* (685), *Staphylococcus capitis* (328), *Campylobacter jejuni* (299)
oxacillin	10985	0.422	*Staphylococcus epidermidis* (4664), *Staphylococcus aureus* (3790), *Staphylococcus hominis* (685), *Staphylococcus capitis* (365), *Staphylococcus lugdunensis* (313)
rifampicin	10966	0.049	*Staphylococcus epidermidis* (4758), *Staphylococcus aureus* (3774), *Staphylococcus hominis* (717), *Staphylococcus capitis* (365), *Staphylococcus lugdunensis* (311)
fusidic acid	10637	0.321	*Staphylococcus epidermidis* (4589), *Staphylococcus aureus* (3766), *Staphylococcus hominis* (692), *Staphylococcus capitis* (365), *Staphylococcus lugdunensis* (311)
gentamicin	10579	0.218	*Staphylococcus epidermidis* (4221), *Staphylococcus aureus* (3629), *Staphylococcus hominis* (683), *Staphylococcus capitis* (319), *Staphylococcus lugdunensis* (297)
cefuroxime	10578	0.423	*Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3640), *Staphylococcus hominis** (668), *Haemophilus influenzae* (347), *Staphylococcus capitis** (322)

antibiotic	number of samples	positive sample ratio	most frequent species
cefazolin	10036	0.421	*Staphylococcus epidermidis** (4261), *Staphylococcus aureus** (3640), *Staphylococcus hominis** (668), *Staphylococcus capitis** (322), *Staphylococcus lugdunensis** (298)
tetracycline	9918	0.311	*Staphylococcus epidermidis* (4132), *Staphylococcus aureus* (3609), *Staphylococcus hominis* (670), *Staphylococcus capitis* (318), *Staphylococcus lugdunensis* (297)
teicoplanin	7691	0.029	*Staphylococcus aureus* (3629), *Enterococcus faecium* (1124), *Enterococcus faecalis* (757), *Staphylococcus epidermidis* (488), *Staphylococcus capitis* (317)
cefpodoxime	6720	0.348	*Escherichia coli* (2075), *Klebsiella pneumoniae* (1826), *Proteus mirabilis* (470), *Enterobacter cloacae* (450), *Klebsiella variicola* (354)
fosfomycin-trometamol	6129	0.216	*Klebsiella pneumoniae* (1809), *Escherichia coli* (1499), *Proteus mirabilis* (465), *Enterobacter cloacae* (443), *Klebsiella variicola* (356)
norfloxacin	6105	0.143	*Klebsiella pneumoniae* (1814), *Escherichia coli* (1488), *Proteus mirabilis* (461), *Enterobacter cloacae* (443), *Klebsiella variicola* (356)
mupirocin	3815	0.007	*Staphylococcus aureus* (3633), *Staphylococcus epidermidis* (84), *MIX!Staphylococcus aureus* (31), *Staphylococcus capitis* (14), *Staphylococcus haemolyticus* (13)
nitrofurantoin	2108	0.195	*Escherichia coli* (1498), *Enterococcus faecium* (464), *Enterococcus faecalis* (54), *Citrobacter freundii* (11), *MIX!Escherichia coli* (10)
aztreonam	856	0.706	*Pseudomonas aeruginosa* (763), *Pseudomonas stutzeri* (22), *Escherichia coli* (13), *Pseudomonas monteilii* (9), *Klebsiella pneumoniae* (8)
gentamicin_high_level	686	0.155	*Enterococcus faecium* (273), *Enterococcus faecalis* (152), *Streptococcus oralis* (43), *Streptococcus mitis* (34), *Streptococcus parasanguinis* (33)
caspofungin	686	0.054	*Candida albicans* (292), *Candida glabrata* (171), *Candida parapsilosis* (65), *Candida tropicalis* (50), *Candida dubliniensis* (29)
5-fluorocytosine	680	0.037	*Candida albicans* (292), *Candida glabrata* (174), *Candida parapsilosis* (65), *Candida tropicalis* (50), *Candida dubliniensis* (29)
micafungin	680	0.151	*Candida albicans* (290), *Candida glabrata* (174), *Candida parapsilosis* (65), *Candida tropicalis* (50), *Candida dubliniensis* (29)
anidulafungin	675	0.283	*Candida albicans* (287), *Candida glabrata* (173), *Candida parapsilosis* (65), *Candida tropicalis* (50), *Candida dubliniensis* (29)
fluconazole	675	0.393	*Candida albicans* (283), *Candida glabrata* (174), *Candida parapsilosis* (65), *Candida tropicalis* (49), *Candida dubliniensis* (29)
itraconazole	668	0.377	*Candida albicans* (283), *Candida glabrata* (172), *Candida parapsilosis* (65), *Candida tropicalis* (43), *Candida dubliniensis* (29)
amoxicillin	629	0.216	*Enterococcus faecium* (91), *Haemophilus influenzae* (56), *Enterococcus faecalis* (52), *Actinotignum schaalii* (40), *Haemophilus parainfluenzae* (36)
amphotericin b	616	0	*Candida albicans* (290), *Candida glabrata* (176), *Candida parapsilosis* (65), *Candida tropicalis* (50), *Candida krusei* (11)
voriconazole	502	0.058	*Candida albicans* (286), *Candida parapsilosis* (65), *Candida tropicalis* (43), *Candida dubliniensis* (27), *Saccharomyces cerevisiae* (25)
posaconazole	412	0.124	*Candida albicans* (282), *Candida parapsilosis* (65), *Candida tropicalis* (43), *Candida glabrata* (4), *Candida albicans_(africana)**** (4)
moxifloxacin	411	0.221	*Finegoldia magna* (23), *Bacteroides fragilis* (19), *Parvimonas micra* (15), *Streptococcus anginosus* (15), *Lactobacillus rhamnosus* (15)
penicillin_with_endokarditis*****	330	0.539	*Streptococcus oralis* (65), *Streptococcus parasanguinis* (56), *Streptococcus mitis* (39), *Streptococcus sanguinis* (23), *Streptococcus gordonii* (19)
penicillin_without_endokarditis*****	325	0.434	*Streptococcus oralis* (65), *Streptococcus parasanguinis* (56), *Streptococcus mitis* (39), *Streptococcus sanguinis* (20), *Streptococcus gordonii* (19)
clarithromycin	313	0.188	*Streptococcus pneumoniae* (255), *Streptococcus agalactiae* (7), *Streptococcus dysgalactiae* (7), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7)
penicillin_with_pneumonia*****	293	0.041	*Streptococcus pneumoniae* (256), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7), *Streptococcus mitis* (6), *Streptococcus constellatus* (5)
penicillin_with_other_infections*****	292	0.175	*Streptococcus pneumoniae* (256), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7), *Streptococcus mitis* (6), *Streptococcus constellatus* (5)
penicillin_with_meningitis*****	291	0.199	*Streptococcus pneumoniae* (255), *Streptococcus pseudopneumoniae* (7), *Streptococcus anginosus* (7), *Streptococcus mitis* (6), *Streptococcus constellatus* (5)

antibiotic	number of samples	positive sample ratio	most frequent species
metronidazole	247	0.057	*Bacteroides fragilis* (37), *Clostridium difficile* (18), *Clostridium perfringens* (13), *Parvimonas micra* (12), *Bacteroides faecis* (9)
quinolones	211	0.057	*Haemophilus influenzae* (55), *Staphylococcus aureus* (19), *Staphylococcus epidermidis* (17), *Corynebacterium macginleyi* (14), *Escherichia coli* (14)
meropenem_with_meningitis*****	210	0.133	*Haemophilus influenzae* (70), *Streptococcus pneumoniae* (56), *Haemophilus parainfluenzae* (40), *MIX!Haemophilus parainfluenzae* (11), *MIX!Haemophilus influenzae* (9)
chloramphenicol	203	0.138	*Haemophilus influenzae* (55), *Staphylococcus aureus* (19), *Staphylococcus epidermidis* (17), *Corynebacterium macginleyi* (14), *Escherichia coli* (14)
meropenem_without_meningitis*****	142	0.035	*Haemophilus influenzae* (72), *Haemophilus parainfluenzae* (40), *MIX!Haemophilus parainfluenzae* (11), *MIX!Haemophilus influenzae* (9), *Haemophilus pittmaniae* (4)
aminoglycosides	122	0.164	*Staphylococcus aureus* (19), *Staphylococcus epidermidis* (15), *Corynebacterium macginleyi* (14), *Escherichia coli* (14), *Pseudomonas aeruginosa* (11)
doxycycline	105	0.114	*Staphylococcus epidermidis* (37), *Finegoldia magna* (6), *Campylobacter jejuni* (5), *Corynebacterium amycolatum* (4), *Haemophilus influenzae* (4)
azithromycin	95	0.126	*Neisseria gonorrhoeae* (67), *MIX!Neisseria gonorrhoeae* (6), *Campylobacter fetus* (5), *Campylobacter coli* (5), *Salmonella spp* (3)
fosfomycin	81	0.58	*Pseudomonas aeruginosa* (37), *Staphylococcus epidermidis* (14), *Klebsiella pneumoniae* (7), *Escherichia coli* (4), *Acinetobacter baumannii* (3)
amoxicillin-clavulanic acid_uncomplicated_UTI*****	80	0.3	*Escherichia coli* (47), *Citrobacter freundii* (14), *Proteus mirabilis* (9), *Klebsiella pneumoniae* (6), *Enterobacter cloacae* (2)
cefixime	74	0.081	*Neisseria gonorrhoeae* (63), *MIX!Neisseria gonorrhoeae* (6), *Neisseria meningitidis* (3), *Yersinia enterocolitica* (2)
minocycline	74	0.297	*Staphylococcus epidermidis* (25), *Burkholderia multivorans* (10), *Stenotrophomonas maltophilia* (7), *Acinetobacter baumannii* (6), *Burkholderia vietnamiensis* (6)
meropenem_with_pneumonia*****	70	0	*Streptococcus pneumoniae* (56), *Streptococcus pseudopneumoniae* (3), *Streptococcus anginosus* (3), *Streptococcus mitis* (3), *Streptococcus constellatus* (3)
cefoxitin_screen	52	0.096	*Staphylococcus aureus* (23), *Staphylococcus epidermidis* (13), *Staphylococcus pseudintermedius* (8), *Staphylococcus hominis* (3), *Staphylococcus lugdunensis* (3)
teicoplanin_grd+	12	0.167	*Staphylococcus aureus* (12)
vancomycin_grd+	12	0	*Staphylococcus aureus* (12)
rifampicin_1mg-l++	10	0.2	*Peptoniphilus harei* (3), *Corynebacterium jeikeium* (2), *Staphylococcus aureus* (2), *Streptococcus oralis* (2), *Corynebacterium simulans* (1)

Supplemental Table 2: Complete list of antibiotics contained in *DRIAMS-A*, including (i) the number of MALDI-TOF mass spectra for which the antimicrobial resistance was determined, (ii) the resistance class ratio among these samples (positive sample ratio) and (iii) the five most frequent species among these samples with the number of samples indicated in brackets. * Antibiotic resistance category is inferred from oxacillin resistance. ** Antibiotic resistance category is inferred from ciprofloxacin resistance. *** Antibiotic resistance category for this species is automatically set to the resistant category. **** Part of a species complex. ***** Breakpoints different according to clinical presentation (see EUCAST guidelines). + vancomycin_grd and teicoplanin_grd is a MIC Strip Test used at University Hospital Basel in very rare cases to detect glycopeptide intermediate *S. aureus*; ++ '1mg_l' indicates the concentration of rifampicin, when MIC are measured in liquid culture as it is routinely done for *Mycobacterium tuberculosis* (MTB). These entries of non-MTB species were entered incorrectly into the laboratory information system.

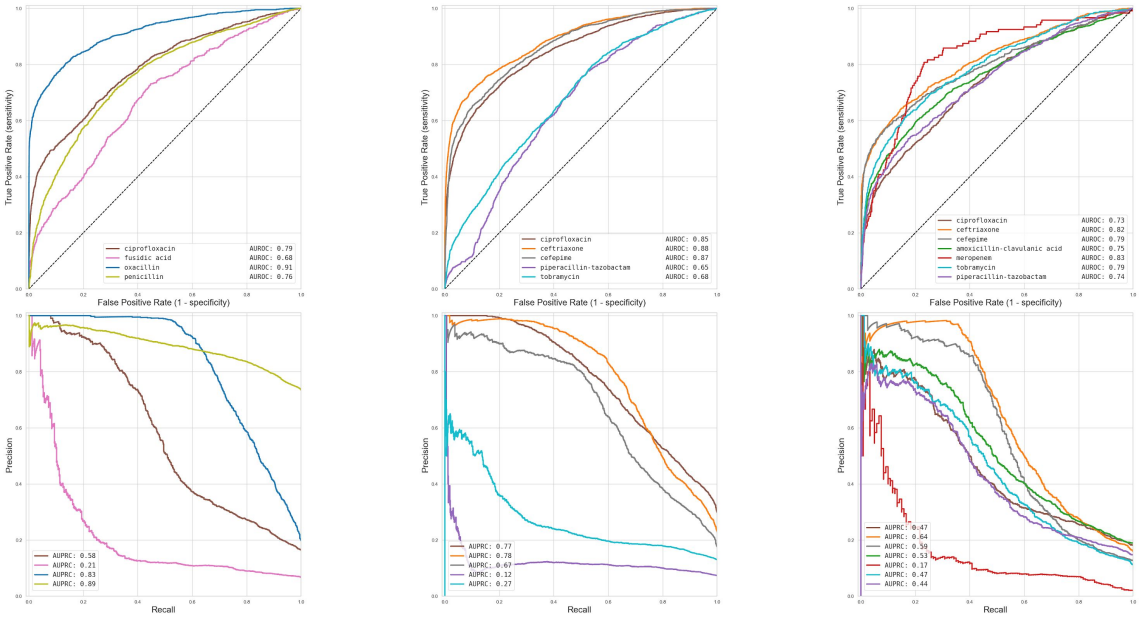
species	antibiotic	model	keep scores less than and assign class 0	keep scores larger than and assign class 1	percentage rejected samples	specificity	sensitivity	TP	FP	TN	FN
E.coli	ceftriaxone	logistic regression	0.025	0.5	75	0.78	0.99	1171	286	1028	13
			0.05	0.5	60	0.9	0.95	1171	286	2465	64
			0.075	0.5	46	0.93	0.88	1171	286	3793	159
			0.1	0.5	36	0.94	0.83	1171	286	4704	241
			0.125	0.5	28	0.95	0.79	1171	286	5356	314
			0.15	0.5	23	0.95	0.75	1171	286	5827	386
S.aureus	oxacillin	logistic regression	0.025	0.5	61	0.93	0.96	864	144	1899	34
			0.05	0.5	42	0.96	0.91	864	144	3276	85
			0.075	0.5	30	0.97	0.86	864	144	4122	140
			0.1	0.5	24	0.97	0.82	864	144	4597	188
			0.125	0.5	19	0.97	0.79	864	144	4873	231
			0.15	0.5	16	0.97	0.76	864	144	5101	271
K. pneumoniae	ceftriaxone	logistic regression	0.025	0.5	79	0.88	0.95	411	88	654	22
			0.05	0.5	54	0.96	0.82	411	88	2041	89
			0.075	0.5	34	0.97	0.72	411	88	3107	156
			0.1	0.5	25	0.98	0.66	411	88	3582	214
			0.125	0.5	19	0.98	0.62	411	88	3890	252
			0.15	0.5	15	0.98	0.6	411	88	4091	278

Supplemental Table 3: Rejection scenarios of calibrated LR classifiers *DRIAMS-A* trained classifiers were calibrated, such that their classification scores become probabilities, resulting in different effects on different models. We simulate scenarios in which samples are rejected if their classification probability falls below a certain threshold. The stricter such a threshold, the more samples are being rejected such that only samples with a high predictive probability are being kept. This can be seen as a 'trade-off' analysis, where strict thresholds lead to better prediction for *some* samples, at the expense of a higher rate of samples that are rejected altogether, i.e. not classified at all. The table illustrates the trade-off between different rejection thresholds for both classes and the amount of rejected/unclassified samples, and how it affects specificity and sensitivity performance. Setting the rejection threshold for the negative (susceptible) class very strictly leads to high performance in sensitivity; the strictest thresholds cause a high percentage of rejected samples and a drop in specificity. However, there is a threshold which keeps both specificity and sensitivity performing well. The reported values depict a class rejection threshold of 0.5, which implies no rejection of positive classified samples. The classifiers are very accurate in positive class prediction and setting a rejection threshold for positive samples does not improve predictive performance drastically. Results with a LightGBM model can be found in **Tab. 1** in the main text.

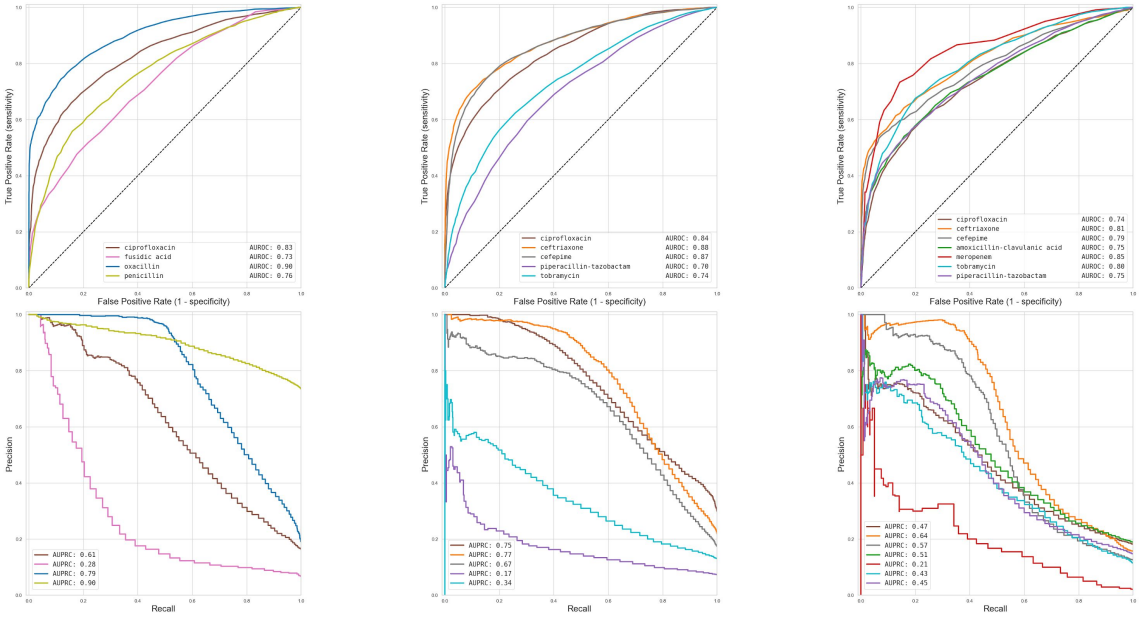
Supplemental Table 4:

A. *Rejection_table_assymetric_Klebsiella pneumoniae_Ceftriaxone_calibrated_lightgbm.csv*
B. *Rejection_table_assymetric_Escherichia coli_Ceftriaxone_calibrated_lightgbm.csv*
C. *Rejection_table_assymetric_Staphylococcus aureus_Oxacillin_calibrated_lightgbm.csv*
D. *Rejection_table_assymetric_Klebsiella pneumoniae_Ceftriaxone_calibrated_lr.csv*
E. *Rejection_table_assymetric_Escherichia coli_Ceftriaxone_calibrated_lr.csv*
F. *Rejection_table_assymetric_Staphylococcus aureus_Oxacillin_calibrated_lr.csv*

A. *S. aureus* (LightGBM) **B.** *E. coli* (LightGBM) **C.** *K. pneumoniae* (LightGBM)

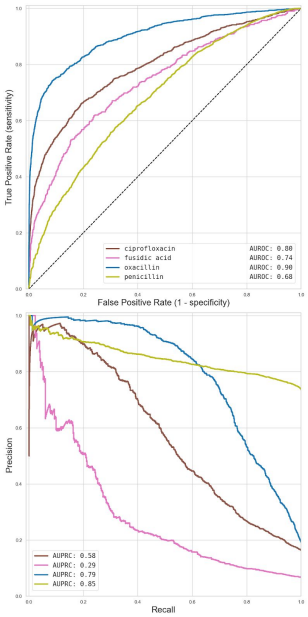


D. *S. aureus* (random forest) **E.** *E. coli* (random forest) **F.** *K. pneumoniae* (random forest)

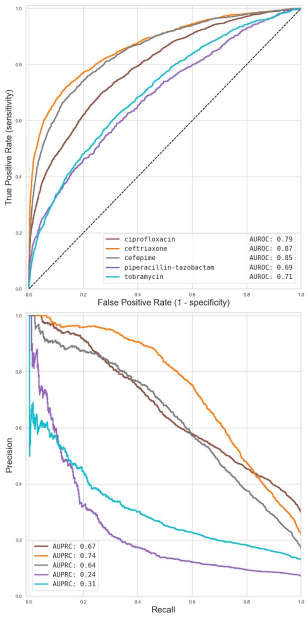


Supplemental Figure 1A: AUROC and AUPRC curves for LightGBM and random forest For each antibiotic, a *DRIAMS-A* subset was created. Each antibiotic dataset from *DRIAMS-A* was divided into stratified training and testing data according to species prevalence. We train a predictor for the clinically relevant species *S. aureus, E. coli*, and *K. pneumoniae*; paired with several antibiotics which are considered for infections with these species. Depicted are the receiver operating characteristic (ROC) curve with AUROC and precision-recall curve with AUPRC value.

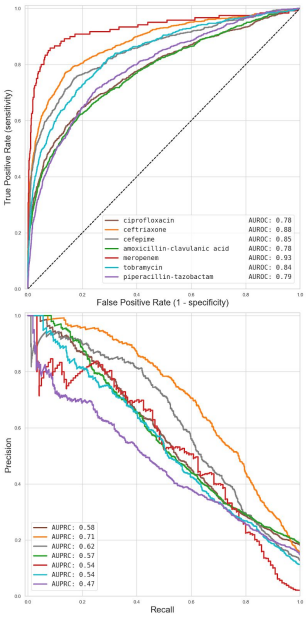
A. *S. aureus* (SVM rbf)



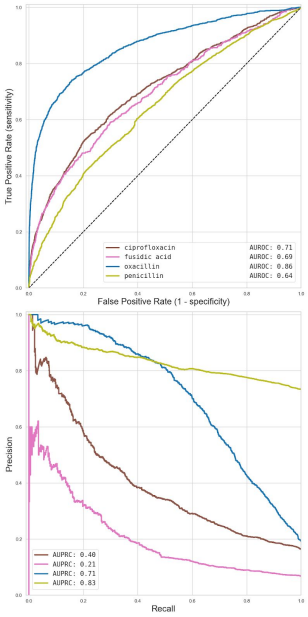
B. *E. coli* (SVM rbf)



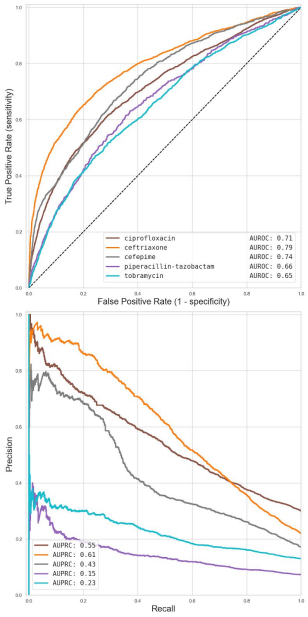
C. *K. pneumoniae* (SVM rbf)



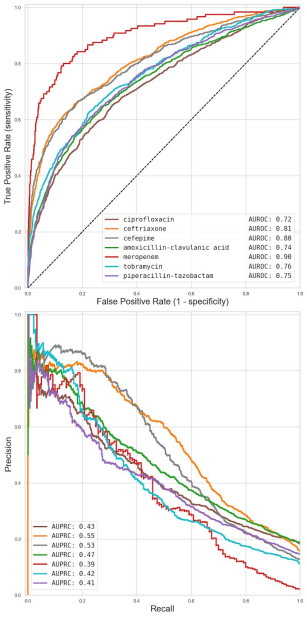
D. *S. aureus* (SVM linear)



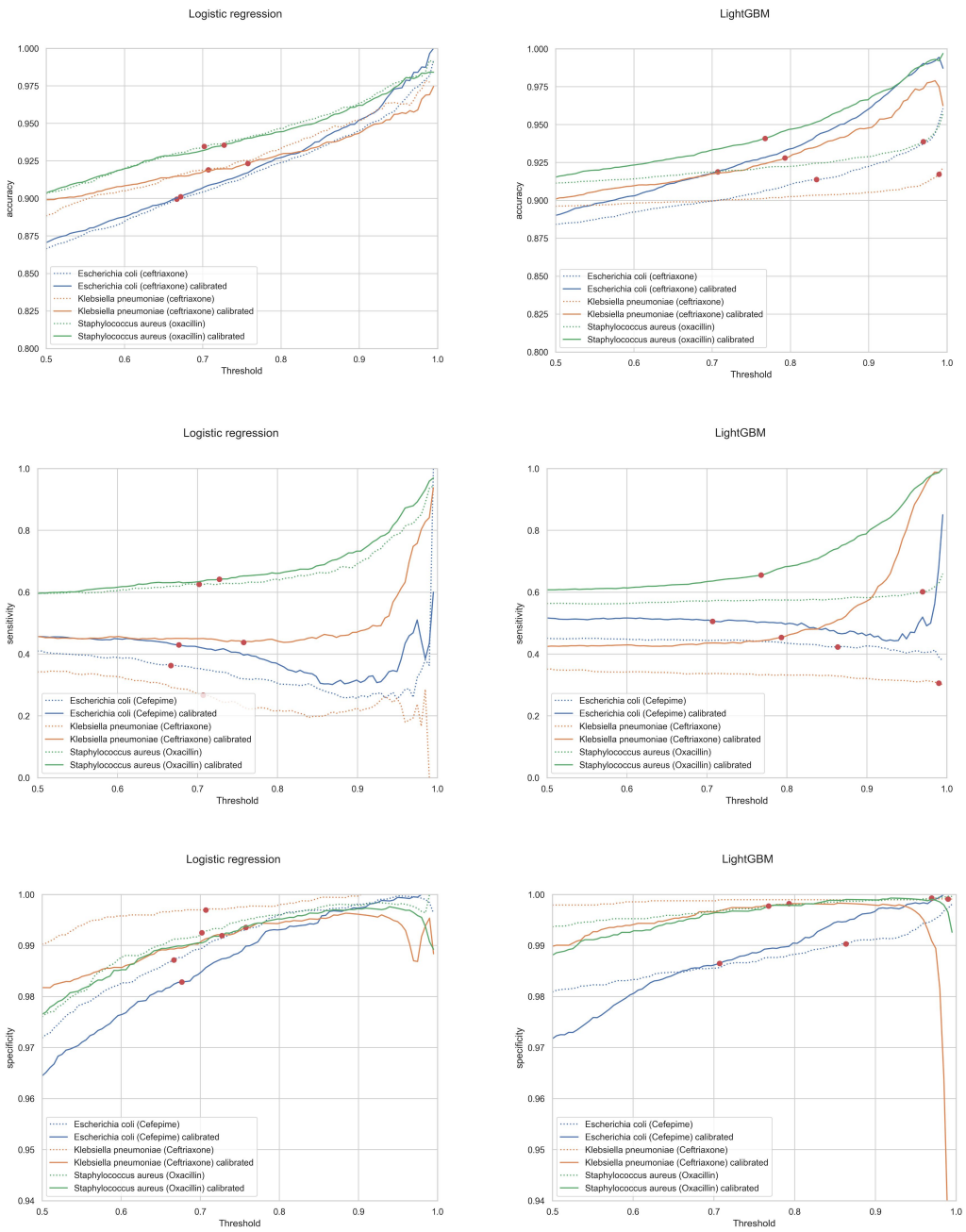
E. *E. coli* (SVM linear)



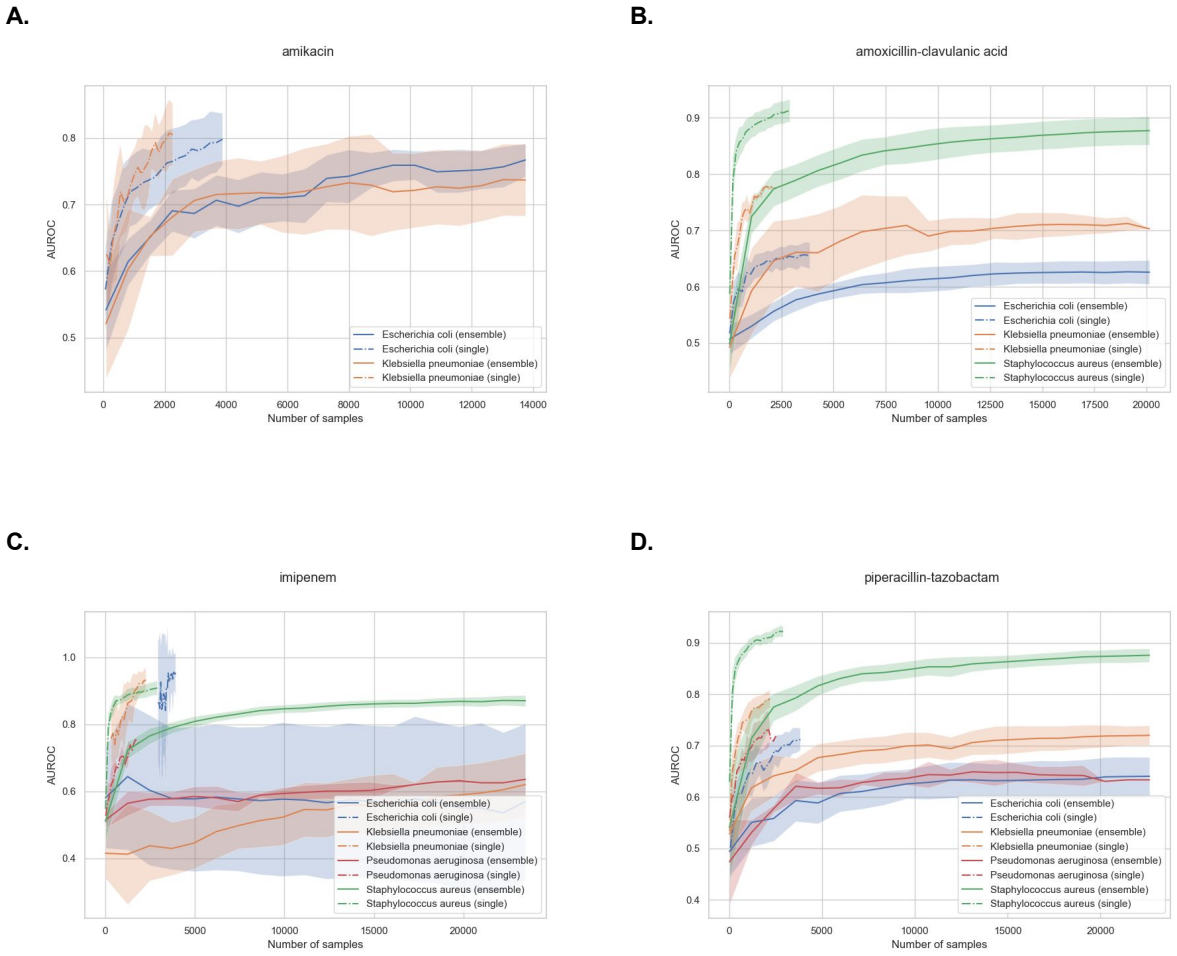
F. *K. pneumoniae* (SVM linear)



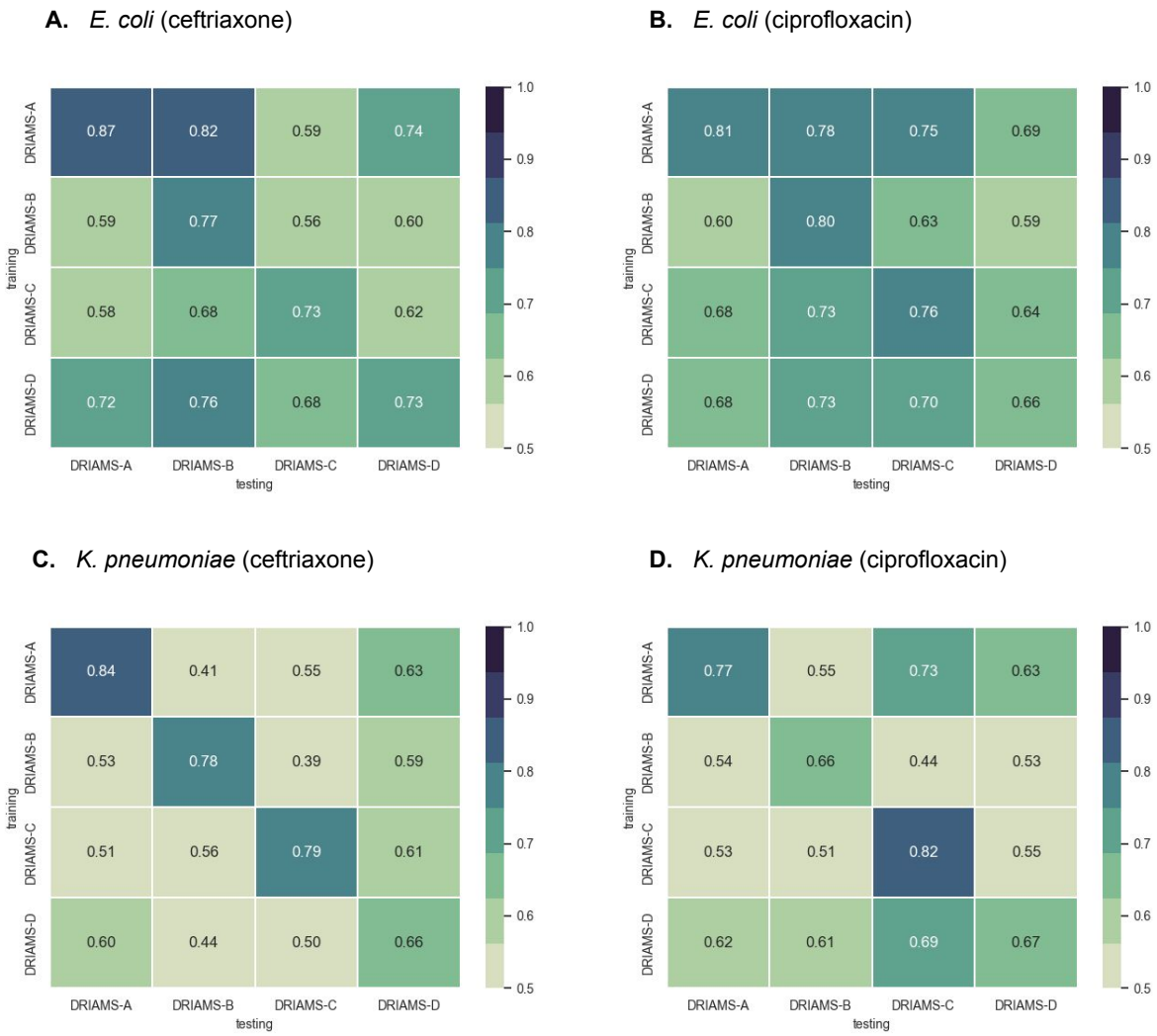
Supplemental Figure 1B: AUROC and AUPRC curves for SVM with linear and rbf kernel For each antibiotic, a *DRIAMS-A* subset was created. Each antibiotic dataset from *DRIAMS-A* was divided into stratified training and testing data according to species prevalence. We train a predictor for the clinically relevant species *S. aureus, E. coli*, and *K. pneumoniae*; paired with several antibiotics which are considered for infections with these species. Depicted are the receiver operating characteristic (ROC) curve with AUROC and precision-recall curve with AUPRC value.



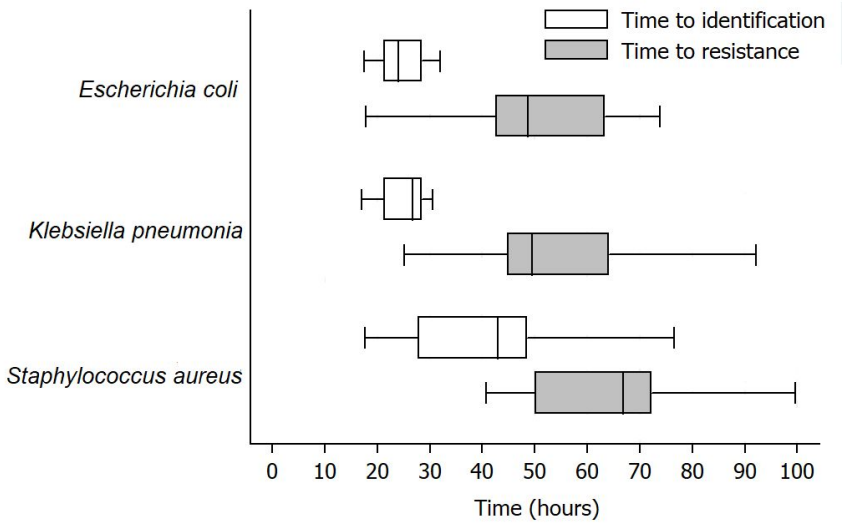
Supplemental Figure 2: Accuracy, sensitivity and specificity for all rejection thresholds applied symmetrically to both positive and negative classes. In the symmetric rejection scenario, we use the maximum class probability (i.e. the class probability of the assigned class) as rejection threshold. We show the differences in uncalibrated (dashed) and calibrated (solid) classifiers. The red dot indicates a rejection rate of 10% of samples. In almost all cases, we observe higher predictive performance for the calibrated classifier (in terms of accuracy, for example). This effect is more pronounced for LightGBM than for logistic regression, as the latter is typically already well-calibrated. Large fluctuations for high rejection thresholds are due to small sample size effects.



Supplemental Figure 3: AUROC with increasing sample size for complete and species-stratified datasets. Experiments were repeated 10 times with different shuffled train-test splits. The solid curve represents the mean of these repetitions, with the envelope depicting the standard deviation over the repetition experiments. The most frequent species are depicted: *S. aureus, E. coli, K. pneumoniae* and *P. aeruginosa*. The depicted antimicrobials are **A.** amikacin **B.** amoxicillin-clavulanic acid **C.** imipenem **D.** piperacillin/tazobactam. *S. aureus* and *P. aeruginosa* were not considered for the amikacin analysis due to insufficient samples from the positive class. Amoxicillin-clavulanic acid resistance was not measured often enough in *P. aeruginosa* to have enough samples for this analysis.



Supplementary Figure 4: Logistic regression predictive performance trained and tested on *DRIAMS-A* to *-D*. The results show the mean AUROC performance of 10 random train-test splits. For comparability, the train-tests split are kept the same for each of the respective four *DRIAMS* datasets. The values reported on the top-right (both training and testing *DRIAMS-A*) correspond to the values reported in **Fig. 2**. With the exception of *DRIAMS-B E.coli*-ceftriaxone, the highest performance is reached when training is performed on the same site as testing. *DRIAMS-A* and *DRIAMS-D* show the highest transferability of predictive performance, and transferability seems higher in general in *E. coli* compared to *K. pneumoniae*.



Supplementary Figure 5: Time to identification and phenotypic resistance testing on three clinical relevant species: *E. coli* (n=54), *K. pneumoniae* (n=66), and *S. aureus* (n=57). Boxplot shows median and interquartile time ranges in hours. Whiskers indicate 5-95% percentile..