nature portfolio

Corresponding author(s): Caroline Weis
Karsten Borgwardt
Adrian Egli

Last updated by author(s): Oct 28, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MALDI-TOF mass spectra were collected using flexControl Software (Bruker Daltonics flexControl v.3.4) and 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)).
Data analysis	All R and Python scripts can be found at https://github.com/BorgwardtLab/maldi_amr. No commercial software was used to analyze the data. The software and package versions are: R-3.4.1, Python 3.7.7 (incl. sklearn 0.24.2)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets generated during and analyzed during the current study are fully available through the Dryad repository at https://doi.org/10.5061/dryad.bzkh1899q.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We have not specifically determined a sample size, but accessed all available MALDI-TOF mass spectral data from the included health care institutes. More than 300,000 mass spectra were accessed for this study. We have generated specific analysis to show the effect of increasing sample size (see manuscript for more details).
Data exclusions	We excluded measurements for which no respective antimicrobial profile could be obtained. Such sample would not be assigned a label in our analysis. We state the number of excluded spectra in the supplement; the data files list all included spectra.
Replication	We used different repetitions (with different random seeds that were chosen after method development had ceased). All experiments are repeated either 5 or 10 times (depending on the data available in each task) to estimate the generalization error. Moreover, fingerprints of the input files are stored with the output of each experiment (which we provide in the associated repository). We made our code public and documented it in order to ensure/facilitate replicability.
Randomization	We used a stratified randomized train-test split to split samples (using fixed random seeds to ensure reproducibility). The stratification is a custom implementation ensuring both stratification by resistance class while also ensuring that measurements of one hospital stay from a patient are present either only in train or only in test, to prevent information leakage (see manuscript for more details). We controlled for no other covariates. For the training of all classifiers, we used cross-validation.
Blinding	The study data was split into training and validation dataset and additional independent validation dataset accessed from three other healthcare institutions. No information making the patient identifiable is included in the data files; they record the MALDI-TOF MS and resistance testing of microbial isolates together with codes only identifiable from within the protected hospital software system.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging