Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data

Allison L. Hicks[1,*] Nicole Wheeler[2], Leonor Sánchez-Busó[3], Jennifer L. Rakeman[4], Simon R. Harris[3], Yonatan H. Grad[1,5,*]

1 Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115

2 Centre for Genomic Pathogen Surveillance, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

3 Pathogen Genomics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

4 Public Health Laboratory, Division of Disease Control, New York City Department of Health and Mental Hygiene, New York City, New York 10016

5 Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115

* Corresponding authors: allison_hicks@g.harvard.edu, ygrad@hsph.harvard.edu

Abstract:

Prediction of antibiotic resistance phenotypes from whole genome sequencing data by machine learning methods has been proposed as a promising platform for the development of sequence-based diagnostics. However, there has been no systematic evaluation of factors that may influence performance of such models, how they might apply to and vary across clinical populations, and what the implications might be in the clinical setting. Here, we performed a meta-analysis of seven large *Neisseria gonorrhoeae* datasets, as well as *Klebsiella pneumoniae* and *Acinetobacter baumannii* datasets, with whole genome sequence and antibiotic susceptibility phenotypes using set covering machine classification, random forest classification, and random forest regression models to predict resistance phenotypes from genotype. We demonstrate how model performance varies by drug, dataset, resistance metric, accuracy metric, and species, reflecting the complexities of generating clinically relevant conclusions from machine learning-derived models. Our findings underscore the importance of incorporating relevant biological and epidemiological knowledge into model design and assessment and suggest that doing so can inform tailored modeling for individual drugs, pathogens, and clinical populations. We further suggest that continued comprehensive sampling and incorporation of up-to-date whole genome sequence data, resistance phenotypes, and treatment outcome data into model training will be crucial to the clinical utility and sustainability of machine learning-based molecular diagnostics.

39 **Introduction:**

40 At least 700,000 deaths annually can be attributed to antimicrobial resistant (AMR)
41 infections, and, without intervention, the annual AMR-associated mortality is estimated to
42 climb to 10 million in the next 35 years[1]. As most patients are still treated based on
43 empirical diagnosis rather than confirmation of the causal agent or its drug susceptibility
44 profile, development of improved, rapid diagnostics enabling tailored therapy represents
45 a clear actionable intervention[1]. The Cepheid GeneXpert MTB/RIF assay, for example,
46 has been widely adopted for rapid point-of-care detection of *Mycobacterium tuberculosis*
47 (TB) and rifampicin (RIF) resistance[2], and the SpeeDx ResistancePlus GC assay used to
48 detect both *Neisseria gonorrhoeae* and ciprofloxacin (CIP) susceptibility was recently
49 approved for marketing as an *in vitro* diagnostic in Europe.

50 Molecular assays offer improved speed compared to gold-standard phenotypic
51 tests and are of particular interest because of their promise of high accuracy for the
52 prediction of AMR phenotype based on genotype[2,3]. Approaches for predicting resistance
53 phenotypes from genetic features include direct association (i.e., using the presence or
54 absence of genetic variants known to be associated with resistance to infer a resistance
55 phenotype) and the application of predictive models derived from machine learning (ML)
56 algorithms. Direct association approaches can offer simple, inexpensive, and often highly
57 accurate resistance assays for some drugs/species[2] and may even provide more reliable
58 predictions of resistance phenotype than phenotypic testing[4-6]. However, these
59 approaches are limited by the availability of well-curated and up-to-date panels of
60 resistance variants, as well as the diversity and complexity of resistance mechanisms.
61 ML strategies can facilitate modeling of more complex, diverse, and/or under-

62 characterized resistance mechanisms, thus outperforming direct association for many
63 drugs/species[7-9]. With the increasing speed and decreasing cost of sequencing and
64 computation, ML approaches can be applied to genome-wide feature sets[8,10-18], ideally
65 obviating the need for comprehensive *a priori* knowledge of resistance loci.

66 While prediction of antibiotic resistance phenotypes from ML-derived models
67 based on genomic features has become increasingly prominent as a promising diagnostic
68 tool[8,11-15,17], there has been no systematic evaluation of factors that may influence
69 performance of such models and their implications in the clinical setting. The extent to
70 which ML model accuracy varies by antibiotic is unclear, as is the impact of sampling bias
71 on model performance. It is further unclear what the most relevant resistance metric (i.e.,
72 minimum inhibitory concentration [MIC] or categorical report of susceptibility) for such a
73 diagnostic might be, how models derived from different methods should be evaluated,
74 and how amenable different species might be to genotype-to-phenotype modeling of
75 antibiotic resistance.

76 We used set covering machine (SCM)[19] and random forest (RF)[20] classification as
77 well as RF regression algorithms to build and test predictive models with seven
78 gonococcal datasets for which whole genome sequences (WGS) and ciprofloxacin (CIP)
79 and azithromycin (AZM) MICs were available. AZM is currently part of the recommended
80 treatment regimen for gonococcal infections, and with the development of resistance
81 diagnostics, CIP may represent a viable treatment option[21-23]. While the majority of CIP
82 resistance in gonococci can be attributed to *gyrA* mutations, AZM resistance is associated
83 with more diverse and complex resistance mechanisms[23,24], offering an opportunity to
84 evaluate ML methods across drugs with distinct pathways to resistance. The range of

85 datasets and sampling frames enables assessment of sampling bias on model reliability.
86 Further, the availability of MICs, as well as distinct EUCAST and CLSI breakpoints, for
87 these drugs allows for evaluation of predictive models based on different resistance
88 metrics and of the implications of different model performance metrics in the clinical
89 setting. Finally, extension of these analyses to *Klebsiella pneumoniae* and *Acinetobacter*
90 *baumannii* datasets for which WGS and CIP MICs were available allows for assessment
91 of model performance for the same drug in species with open pangenomes.

92 Our results demonstrate that using ML to predict antibiotic resistance phenotypes
93 from WGS data yields variable results across drugs, datasets, resistance metrics, metrics
94 of model performance, and species. Ultimately, we suggest that tailored modeling for
95 individual drugs, species, and clinical populations may be necessary to successfully
96 leverage these ML-based approaches as diagnostic tools. We further suggest that
97 continuing surveillance, isolate collection, and reporting of WGS, MIC phenotypes, and
98 treatment outcomes will be crucial to the sustainability of any such molecular diagnostics.
99
100

101 **Methods:**

102 **Isolate selection and dataset preparation**

103 See **Table 1** for details of the datasets assessed. All gonococcal datasets contained a
104 minimum of 200 isolates with WGS (Illumina MiSeq, HiSeq, or NextSeq) and MICs
105 available for both CIP and AZM (by agar dilution and/or Etest). Isolates lacking CIP and
106 AZM MIC data were excluded. MIC testing methods varied within datasets, as reported[10-
107 13,17,18,25].

108 *K. pneumoniae* and *A. baumannii* datasets were selected based on the availability
109 of isolates collected during a single survey that were tested for CIP susceptibility and
110 whole genome sequenced using consistent platforms (in both cases, the BD-Phoenix
111 system and either Illumina MiSeq or NextSeq).

112 MIC data were obtained from the associated publications, except in the cases of
113 dataset 1 (NCBI Bioproject PRJEB10016; see **Supplementary Table 1**) and dataset 9,
114 which were obtained from the NCBI BioSample database
115 (https://www.ncbi.nlm.nih.gov/biosample). Raw sequence data were downloaded from
116 the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra). Genomes were
117 assembled using SPAdes[26] with default parameters, and assembly quality was assessed
118 using QUAST[27]. Contigs <200 bp in length and/or with <10x coverage were removed.
119 Isolates with assembly N50s below two standard deviations of the dataset mean were
120 removed.

121

122 **Evaluation of known resistance variants**

123 Previously identified genetic loci associated with reduced susceptibility to CIP or AZM in
124 gonococci are indicated in **Supplementary Tables 2** and **3**, respectively. The sequences
125 of these loci were extracted from the gonococcus genome assemblies using BLAST[28]
126 followed by MUSCLE alignment [29] to assess the presence or absence of known
127 resistance variants. The presence or absence of quinolone resistance determining
128 mutations in *gyrA* was similarly assessed in *K. pneumoniae* and *A. baumannii*
129 assemblies. Presence or absence of gonococcal AZM resistance mutations in the multi-
130 copy 23S rRNA gene was assessed using BWA-MEM[30] to map raw reads to a single 23S
131 rRNA allele from the NCCP11945 reference isolate (NGK_rrna23s4), the Picard toolkit

132 (http://broadinstitute.github.io/picard) to identify duplicate reads, and Pilon[31] to determine
133 the mapping quality-weighted percentage of each nucleotide at the sites of interest.
134
135 **ML-based prediction of resistance phenotypes**
136 Predictive modeling was carried out using SCM and RF algorithms, implemented in the
137 Kover[11,12] and ranger[32] packages, respectively. K-mer profiles used for model training
138 and prediction were generated from the assemblies using the DSK k-mer counting
139 software[33] with k=31, a length commonly used in bacterial genomic analysis[11,12,34,35]. For
140 each SCM binary classification analysis (using S/NS phenotypes based on the two
141 different breakpoints for each drug), the best conjunctive and/or disjunctive model was
142 selected using five-fold cross-validation, testing the suggested broad range of values for
143 the trade-off hyperparameter of 0.1, 0.178, 0.316, 0.562, 1.0, 1.778, 3.162, 5.623, 10.0,
144 and 999999.0 to determine the optimal rule scoring function
145 (http://aldro61.github.io/kover/doc_learning.html) with default parameters. In order to
146 assess binary classification across multiple methods, RF was also used to build binary
147 classifiers (RF-C) using S/NS phenotypes. Further, to compare performance of binary
148 classifiers to MIC prediction models, RF was used to build multi-class classification (RF-
149 mC) and regression (RF-R) models based on $\log_2$(MIC) data. For all RF analyses, forests
150 were grown to 1000 trees of unlimited depth using node impurity to assess variable
151 importance using default parameters.
152 The set of SCM and RF analyses performed are indicated in **Supplementary**
153 **Tables 4** and **5.** For each of the seven individual gonococcal datasets, as well as the
154 aggregate gonococcal dataset and the *K. pneumoniae* and *A. baumannii* datasets,

155 training sets consisted of random sub-samples of two-thirds of isolates from the dataset
156 indicated (maintaining proportions of each resistance phenotype from the original
157 dataset), while the remaining isolates were used to test performance of the model. Each
158 set of analyses (for each combination of dataset/drug/resistance metric/ML algorithm)
159 was performed on 10 replicates, each with a unique randomly partitioned training and
160 testing set. For all gonococcal datasets, separate models were trained and tested using
161 the EUCAST[36] and CLSI[37] breakpoints for non-susceptibility (NS) to CIP. Four of the *N.*
162 *gonorrhoeae* datasets had insufficient (<15) NS isolates by the CLSI breakpoint for AZM
163 non-susceptibility[37] and thus were only assessed at the EUCAST AZM breakpoint. CIP
164 MICs for the *K. pneumoniae* isolates were not available in the range of the EUCAST
165 breakpoint (0.25 μg/mL), and thus only the CLSI breakpoint for NS was assessed. For *A.*
166 *baumannii*, the EUCAST and CLSI breakpoints for ciprofloxacin NS are the same (>1
167 μg/mL). Due to the very limited range of MICs within the BD-Phoenix testing thresholds
168 and thus the CIP MICs available for *K. pneumoniae* and *A. baumannii*, predictive models
169 based on MICs were not generated for these species.

170 Model performance was assessed by sensitivity (1 – very major error [VME] rate),
171 specificity (1 – major error [ME] rate), and the aggregate balanced accuracy (bACC). For
172 MIC prediction models, the percentage of isolates with predicted MICs exactly matching
173 the phenotypic MICs (rounding to the nearest doubling dilution, in the case of regression
174 models), as well as the percentage of isolates with predicted MICs within one doubling
175 dilution of phenotypic MICs (1-tier accuracy), were also assessed. Mean and 95%
176 confidence intervals for all metrics were calculated across the 10 replicates for each
177 analysis. Differential model performance between datasets or methods was evaluated by

178 comparing mean bACC between sets of replicates by two-tailed t-tests (α=0.05).
179 Relationships between MIC prediction accuracy and bACC and between dataset
180 imbalance and model performance were assessed by Pearson correlation (α=0.05).
181
182 **Results:**
183 **Accuracy of ML-based prediction of resistance phenotypes varies by antibiotic.**
184 Given the distinct MIC distributions and distinct pathways to resistance for CIP and AZM
185 in gonococci, these two drugs enable evaluation of drug-specific performance of ML-
186 based resistance prediction models. CIP MICs in surveys of clinical gonococcal isolates
187 are bimodally distributed, with the majority of isolates having MICs well above or below
188 the NS breakpoints, while the majority of reported AZM MICs in gonococci are closer to
189 the NS breakpoints (https://mic.eucast.org/Eucast2). These trends were recapitulated in
190 the gonococcal isolates assessed here (**Fig. 1a-b**). Further, the vast majority of CIP
191 resistance in gonococci observed to date is explained by mutations in *gyrA* and *parC* and
192 has spread predominantly through clonal expansion, generally resulting in MICs ≥ 1
193 μg/mL[23,38]. In contrast, AZM resistance in gonococci has arisen many times *de novo*
194 through multiple pathways, many of which remain under-characterized and are
195 associated with lower-level resistance[23,38,39]. As expected, the GyrA S91F mutation alone
196 predicts NS to CIP by both EUCAST and CLSI breakpoints in the aggregate gonococcal
197 dataset assessed here with ≥98% sensitivity and ≥99% specificity (**Supplementary Table**
198 **2**). AZM NS showed lower values for these metrics, indicating it was not as well explained
199 by known resistance variants, with extensive contributions from uncharacterized
200 mechanisms and/or multifactorial interactions (**Supplementary Table 3**).

201 We next trained and evaluated ML-based predictive models for CIP and AZM
202 resistance in gonococci (**Supplementary Table 4**). By all ML methods and breakpoints,
203 CIP NS was predicted with significantly higher bACC than AZM NS in the aggregate
204 gonococcal dataset ($P < 0.0001$, **Fig. 1c-d**), as well as in individual gonococcal datasets
205 ($P < 0.0001$, **Supplementary Tables 6-7**). While CIP NS was predicted with mean bACC
206 ≥96% across all methods, breakpoints, and datasets, mean bACC for AZM NS
207 classification ranged from 62% to 92%, varying by method, breakpoint, and dataset. As
208 variable model performance across different drugs has previously been attributed to
209 variations in representation of susceptible (S) or NS isolates[7,14,15], it is worth nothing that
210 by the EUCAST breakpoints, the aggregate gonococcal dataset, as well as some of the
211 individual datasets, had nearly identical proportions of S and NS isolates between CIP
212 and AZM, demonstrating that variable representation of S or NS isolates alone cannot
213 explain reduced performance of AZM models compared to CIP.
214
215 **Sampling bias in training and testing data skews resistance model performance.**
216 The diversity of resistance mechanisms for AZM in gonococci offers an opportunity to
217 evaluate the effects of sampling bias on model performance. The sampling frames for the
218 seven gonococcal datasets ranged geographically from citywide to international and
219 temporally from a single year to >20 years (**Table 1**), and several datasets were enriched
220 for AZM resistance[11,25]. The distributions of both AZM MICs and known resistance
221 mechanisms across datasets (**Fig. 1b, Supplementary Table 3**) and the variable
222 performance of AZM resistance models across datasets (**Supplementary Table 7**)
223 suggest that AZM resistance mechanisms are differentially distributed across the

224 sampled clinical populations. To assess the impact of sampling on model reliability, the
225 performance of RF classifiers in prediction of AZM NS phenotypes were compared across
226 multiple training and testing sets. These include classifiers trained on subsamples of
227 isolates from a single dataset, classifiers trained on the aggregate gonococcal dataset,
228 and classifiers trained on the aggregate gonococcal dataset excluding isolates from the
229 same dataset as the testing set (**Supplementary Table 5**). Given the low representation
230 of AZM NS strains by the CLSI breakpoint in many datasets, these analyses were only
231 performed using the EUCAST breakpoint.

232 While it may be assumed that increased availability of paired genomic and
233 phenotypic resistance data from a broader range of clinical populations will facilitate more
234 accurate and reliable modeling[40], our results demonstrate that in predicting AZM
235 resistance phenotypes for isolates from most datasets (with the exception of datasets 2
236 and 5), performance of classifiers trained on the aggregate dataset was not significantly
237 better than performance of classifiers trained only on isolates from the dataset from which
238 the test isolates were derived ($P < 0.0001$ and $P = 0.002$ for datasets 2 and 5,
239 respectively, $P = 0.019$ for dataset 3, where the classifiers trains on the aggregate dataset
240 had lower bACC than classifiers trained only on isolates from dataset 3, and $P > 0.25$ for
241 all other datasets, **Fig. 2a**). Further, there was substantial variation in performance of
242 models trained on the aggregate dataset across testing sets, with models achieving
243 significantly higher bACC for strains from datasets 3 and 4 than for strains from datasets
244 2 and 5 ($P < 0.004$, **Fig. 2a**), perhaps reflecting enrichment for AZM NS in these datasets
245 (**Table 1**). Additionally, with the exception of dataset 5, performance of AZM resistance
246 classifiers trained only on isolates from the dataset from which the test isolates were

247 derived was significantly higher than performance of classifiers trained on the aggregate
248 dataset excluding isolates from the test dataset (P = 0.392 for dataset 5, P < 0.01 for all
249 other datasets, **Fig. 2a**).

250 Performance of RF classifiers trained and tested on dataset 2 was limited by low
251 specificity, which was improved in models trained on the aggregate dataset (**Fig. 2b**). The
252 low specificity achieved by RF classifiers trained and tested on this dataset is likely due
253 to the low representation of S strains, most of which were within one doubling dilution of
254 the NS breakpoint (**Fig. 2c**), and thus the more comprehensive representation of negative
255 (S) data in the aggregate training set was associated with improved specificity.
256 Conversely, performance of RF classifiers trained and tested on dataset 5 was more
257 limited by low sensitivity, which was improved in models trained on the aggregate dataset
258 (**Fig. 2b**). This dataset had a low representation of strains with high AZM MICs (**Fig. 2d**),
259 and thus the more comprehensive representation of positive (NS) data in the aggregate
260 training set was associated with improved sensitivity in predicting AZM NS for these
261 strains. Low representation of strains with higher AZM MICs was also observed in other
262 datasets (i.e., datasets 1, 6, and 7) and was similarly reflected in the sensitivity-limited
263 performance of RF classifiers trained and tested on these datasets (**Supplementary**
264 **Table 7**). However, AZM NS prediction accuracy for strains from these datasets was not
265 improved by training classifiers on the aggregate dataset. These results demonstrate that
266 resistance model performance may be strongly associated with the distributions of both
267 resistance phenotypes and genetic features and thus can be highly population-specific.
268

269 **ML prediction models of antibiotic susceptibility / non-susceptibility outperform**
270 **MIC models**

271 Gonococcal CIP and AZM MICs were dichotomized by both EUCAST and CLSI
272 breakpoints to assess the impact of variation in MIC breakpoints on model performance.
273 As the EUCAST and CLSI breakpoints for CIP in gonococci are within a single doubling
274 dilution and the vast majority of isolates have much lower or higher CIP MICs (**Fig. 1a**),
275 >99% of isolates in the aggregate dataset were consistently S or NS by both breakpoints.
276 Of the 23 isolates with MICs between the two breakpoints, 18 had MICs derived from
277 Etests of 0.032 μg/mL or 0.047 μg/mL, making their classification relative to the EUCAST
278 breakpoint of 0.03 μg/mL ambiguous. In contrast, the EUCAST and CLSI breakpoints for
279 AZM in gonococci are separated by two doubling dilutions, and for many isolates, the
280 AZM MIC was within this range (**Fig. 1b**). As such, only 67% of isolates in the aggregate
281 dataset were consistently S or NS by both breakpoints. CIP NS classifier performance
282 was either identical or nearly identical for both breakpoints in the aggregate and most
283 individual gonococcal datasets (**Fig. 3a**). In contrast, the bACC of AZM NS prediction by
284 both SCM and RF classifiers based on the CLSI breakpoint was significantly higher than
285 for those based on the EUCAST breakpoint across all gonococcal datasets assessed by
286 both breakpoints ($P < 0.0001$, **Fig. 3b**).

287 To assess the performance of MIC prediction models relative to binary S/NS
288 resistance phenotype classifiers, RF-mC and RF-R models were trained and evaluated
289 for CIP and AZM MIC prediction in gonococci. Average exact match rates between
290 predicted and phenotypic MICs ranged from 63-86% and 53-77% by RF-mC and RF-R,
291 respectively, for CIP, and from 22-58% and 44-64%, respectively, for AZM

292 (**Supplementary Tables 6-7**). Average 1-tier accuracies were substantially higher but
293 similarly varied widely across datasets and between the two MIC prediction methods.
294 There was no consistent or significant relationship across the different datasets between
295 MIC prediction accuracy (exact match or 1-tier accuracy) and bACC for either drug by
296 either MIC prediction method (**Fig. 3c-f**). Further, for both drugs by both breakpoints in
297 the aggregate gonococcal dataset, binary RF-C models had equivalent or significantly
298 higher bACC than RF-mC and RF-R MIC prediction models (P = 0.513 for CIP NS by the
299 CLSI breakpoint by RF-C compared to RF-R, P = 0.201 for AZM NS by the CLSI
300 breakpoint by RF-C compared to RF-R, P < 0.0006 for all others, **Supplementary Tables**
301 **6-7**).
302
303 **Model performance varies substantially across performance metrics**
304 Success in the predictive accuracy of ML models varies not only by antibiotic, dataset,
305 and ML method, but also by metrics used to assess model performance[7-12,14,15,17,25,41]. To
306 assess the advantages and limitations of model performance metrics and their
307 implications for diagnostics, we examined the performance of predictive models for AZM
308 resistance in gonococci across multiple metrics. Specifically, we evaluated accuracy (1 -
309 error rate) compared to the bACC across all models for AZM S/NS based on the EUCAST
310 breakpoint, and bACC was further compared to individual metrics of sensitivity (1 – VME
311 rate) and specificity (1 – ME rate). Given the low representation of AZM NS strains by the
312 CLSI breakpoint in most datasets, comparison of performance metrics was limited to
313 models based on the EUCAST breakpoint.

314 Model accuracy was significantly higher than bACC for SCM and RF-C AZM
315 resistance models in all gonococcal datasets ($P < 0.0001$), except the aggregate dataset
316 and dataset 6 ($P > 0.40$), with a particularly marked discordance in datasets with
317 unbalanced representation of S and NS phenotypes (**Fig 4a-c**). For example, in dataset
318 2, there were almost 5x as many AZM NS strains as S strains by the EUCAST breakpoint
319 (**Fig. 4a**, **Supplementary Table 7**). While the mean error rate across the SCM replicates
320 for this dataset based on this breakpoint was 15% (accuracy = 85%), this obscures the
321 low specificity, which is better reflected in the mean bACC of 62%. However, even
322 normalized aggregate metrics, such as bACC, can fail to reflect differences in sensitivity
323 vs. specificity across models (**Fig. 4d-e**). For example, models trained and tested on
324 dataset 1 had significantly higher bACC across both ML methods than models from
325 dataset 2 ($P < 0.0001$), while the models from the dataset 2 had 38-47% higher sensitivity.
326 For both SCM and RF-C AZM resistance models, there was a significant positive
327 correlation between the ratio of model sensitivity to model specificity and the ratio of NS
328 to S strains in the dataset (Pearson $r > 0.98$, $P < 0.0001$ for both SCM and RF-C, **Fig.**
329 **4f**).
330

331 **Species with large accessory genomes pose challenges to ML-based antibiotic**
332 **resistance prediction**

333 Increasing pangenome size, or increasing ratio of genomic features to observations, may
334 present an additional challenge for ML-based prediction of antibiotic resistance[12]. To
335 investigate the impact of pangenome size on ML-based antibiotic resistance prediction,
336 SCM and RF-C were used to model CIP NS in *K. pneumoniae* and *A. baumannii*, two

337 species with pangenomes several times that of gonococci (**Fig. 5a-b**). SCM classifiers
338 trained on and used to predict CIP NS for *K. pneumoniae* and *A. baumannii* achieved
339 significantly lower or roughly equivalent accuracy, respectively, as the gonococcal
340 datasets ($P < 0.0001$ and $P > 0.06$ for *K. pneumoniae* and *A. baumannii*, respectively,
341 **Fig. 5c**), and the performance of RF-C models was significantly lower for both *K.*
342 *pneumoniae* and *A. baumannii* ($P < 0.0001$, **Fig. 5d**). Direct association based on GyrA
343 codon 83 mutations (equivalent to codon 91 in gonococci) alone predicted CIP NS in *K.*
344 *pneumoniae* with 86% sensitivity and 99% specificity, and thus had a marginally higher
345 bACC (92.5%) than for the SCM classifiers and a substantially higher bACC than the RF
346 classifiers. Similarly, for *A. baumannii*, GyrA codon 81 mutations (equivalent to codon 91
347 in gonococci) alone predicted CIP NS in with 97% sensitivity and 98% specificity, and
348 thus with a roughly equivalent bACC (97.5%) to the SCM classifiers and a substantially
349 higher bACC than the RF classifiers.
350
351
352 **Discussion**
353 ML offers an opportunity to leverage WGS data to aid in development of rapid molecular
354 diagnostics, but multiple factors affect model performance, reliability, and interpretability.
355 Our results affirmed that drugs associated with complex and/or diverse resistance
356 mechanisms present challenges to ML-based prediction of resistance phenotypes, and
357 sampling frame can substantially affect performance of such predictive models. We
358 demonstrated significant variability in performance and potential clinical utility of
359 predictive models based on different resistance metrics, as well as in the information

360 provided by, and thus the clinical applicability of, commonly used metrics of model
361 performance. We further showed that the capacity to model antibiotic resistance may be
362 highly variable across different species.
363
364 **Variable performance of ML-based resistance prediction models by antibiotic**
365 Genotype-based resistance diagnostics have largely focused more on evaluating
366 the presence of resistance determinants and less on predicting the susceptibility profile
367 of a given isolate[8]. However, in clinical settings where the empirical presumption is of
368 resistance, prediction that an isolate is susceptible to an antibiotic may be more important
369 in guiding treatment decisions. As such, the clinical utility of a genotype-based resistance
370 diagnostic may be determined by its capacity to accurately predict susceptibility
371 phenotype for multiple drugs.
372 While variable performance of ML-based predictive models has been observed
373 across different drugs[7,8,10,11,14,15], it has often been attributed to dataset size and/or
374 imbalance[7,14,15]. Further, while it is more difficult to predict resistance phenotypes from
375 genotypes for drugs that are associated with unknown, multifactorial, and/or diverse
376 resistance mechanisms than for drugs for which resistance can largely be attributed to a
377 single variant[14,25], this caveat has been presented specifically as a limitation of models
378 based on known resistance loci in comparison to unbiased machine learning-based MIC
379 prediction using genome-wide feature sets[14]. However, by comparing performance of
380 predictive models based on genome-wide feature sets between CIP and AZM across
381 multiple gonococcal datasets, we showed that even with relatively large and
382 phenotypically balanced datasets, ML algorithms cannot necessarily be expected to

383 successfully model complex and/or diverse resistance mechanisms, particularly given
384 that the representation of these resistance mechanisms in training datasets is *a priori*
385 unknown.
386
387 **Impact of demographic, geographic, and timeframe sampling bias on ML model**
388 **predictions of antibiotic resistance**
389 Sampling bias presents a substantial challenge in any predictive modeling, and
390 sampling from limited patient demographics or during limited time periods may have
391 considerable effects on the distributions of resistance phenotypes and resistance
392 mechanisms[42,43]. For example, in TB, the RpoB I491F mutation that has been associated
393 with failure of commercial RIF resistance diagnostic assays, including the GeneXpert
394 MTB/RIF assay, reportedly accounted for <5% of TB RIF resistance in most countries,
395 but, in Swaziland was found to be present in up to 30% of MDR-TB[44]. Further, as the
396 focus with statistical classifiers is building models from feature sets that can accurately
397 predict an outcome, rather than understanding the association between each of the
398 features and the outcome, potential confounding effects from factors such as population
399 structure[35,45,46] or correlations among resistance profiles of different drugs[13] are rarely
400 considered.
401 By comparing performance of AZM NS classifiers across multiple training and
402 testing sets, we showed significant variation in performance of classifiers trained on a
403 large and diverse global collection across testing sets from different sampling frames. In
404 some cases of imbalanced datasets, models trained on datasets with a more
405 comprehensive representation of resistance phenotypes improve prediction accuracy.

406 However, our results suggest that heavier sampling across more geographic regions
407 cannot necessarily be expected to significantly improve model performance. This,
408 together with decreased performance when excluding isolates from the dataset from
409 which the isolates being tested were derived, suggests that factors such as population-
410 specific resistance mechanisms, genetic divergence at resistance loci, and/or
411 confounding effects may constrain model reliability across populations.
412

413 **ML resistance prediction model performance varies by NS breakpoints and by**
414 **categorical vs MIC-based resistance metrics**

415 While measurement of MICs is vital for surveillance and investigation of resistance
416 mechanisms, resistance breakpoints that relate *in vitro* MIC measurements to expected
417 treatment outcomes inform clinical decision-making. However, standard breakpoints for
418 NS to a given drug in a given species are often informed less by treatment outcome data,
419 but rather factors such as pharmacokinetics and MIC distributions that can fail to account
420 for a variety of intra-host conditions that could influence drug efficacy[47-50]. Recent studies
421 have shown that isolates that are classified as susceptible by standard breakpoints but
422 have higher MICs are associated with a greater risk of treatment failure than isolates with
423 lower MICs[51]. Further, resistance breakpoints and testing protocols can vary across
424 different organizations, and thus incongruence across phenotypic information included in
425 the training data may introduce additional sources of error in predictive modeling. By
426 comparing performance of predictive models of CIP and AZM non-susceptibility based on
427 EUCAST and CLSI breakpoints, we demonstrated breakpoint-specific performance of
428 models. For CIP, such breakpoint-specific performance is likely largely attributable to

429 variations in MIC testing protocols and thus ambiguous classification of some strains by
430 the EUCAST breakpoint. On the other hand, the substantially lower performance of all
431 AZM models based on the EUCAST breakpoint compared to those based on the CLSI
432 breakpoint suggests that many isolates with AZM MICs between the two breakpoints lack
433 genetic signatures that contribute to high model performance. While the clinical relevance
434 of AZM MICs between these two breakpoints in gonococci is unclear, these isolates may
435 be more likely to be associated with AZM treatment failure than isolates with lower MICs,
436 and thus evaluation of classifiers using only higher breakpoints may misrepresent their
437 diagnostic value, particularly in the absence of sufficient treatment outcome data.

438 Models that predict MICs provide more refined output than a binary classifier but
439 generally achieve low rates of exact matches between phenotypic and predicted MICs
440 and even fairly variable 1-tier accuracies[14,15,25]. Given the noise in phenotypic MIC
441 testing[52] and the potential lack of discriminating genetic features between isolates with
442 MICs separated by 1-2 doubling dilutions[14], MIC prediction models may be unlikely to
443 provide much better resolution than binary S/NS classifiers. Further, even if MIC
444 predictions could provide additional resolution, the most important criterion of such a
445 diagnostic would likely still be its ability to correctly predict resistance phenotypes relative
446 to a clinically relevant breakpoint. Thus, performance of MIC prediction models with
447 respect to breakpoints may be the biggest determinant of their diagnostic utility. By
448 building MIC prediction models for CIP and AZM in gonococci, we observed low rates of
449 exact matches between phenotypic and predicted MICs and variable 1-tier accuracies,
450 with no relationship between 1-tier accuracy and categorical agreement (i.e., prediction

451 accuracy relative to NS breakpoints). Further, binary classifiers performed equivalently or
452 better than MIC prediction models.
453
454 **The choice of model performance metrics can obscure shortcomings of**
455 **resistance prediction models**
456 While performance can vary substantially across resistance prediction models built
457 by different ML methods[7,12,53], criteria for selecting a model with the greatest potential
458 diagnostic value are seldom addressed. Performance assessments for resistance
459 prediction models are frequently presented in terms of aggregate metrics, including
460 accuracy (or error rate), area-under-the-ROC-curve (AUC), and 1-tier accuracy, and/or in
461 terms of individual VME and ME rates (or the sensitivity and specificity, respectively)[7-
462 12,14,15,17,25,41]. Aggregate metrics can be useful in providing a single intuitive measure of
463 model performance. However, as previously noted, metrics such as 1-tier accuracy may
464 not reflect model performance relative to utility as a diagnostic (i.e., what proportion of
465 discrepancies between phenotypic and predicted MICs result in a VME or ME). Further,
466 some of these metrics, such as accuracy (or error rate) and AUC, may provide skewed
467 representations of model performance in the case of imbalanced datasets[54].
468 Comparisons of AZM NS classifier accuracy to bACC across each of the gonococcal
469 datasets demonstrated that accuracy obscures performance deficiencies. However, even
470 normalized aggregate metrics such as bACC can fail to capture potentially important
471 differences in sensitivity vs. specificity (or VME vs. ME rates). Individual metrics of
472 sensitivity and specificity provide more detailed information about the likelihood of
473 different kinds of prediction failures, the differential importance of which is reflected in the

474 FDA guidelines for AMR diagnostics[55]. However, our results also illustrate that model
475 sensitivity and specificity can be strongly influenced by dataset imbalance, ultimately
476 suggesting that multiple metrics may be necessary to evaluate a model's clinical utility
477 and that both comprehensive sampling and dataset pruning may be necessary to optimize
478 model performance.
479
480 **ML antibiotic resistance prediction model success varies by bacterial pangenome**
481 **size**
482 Bacterial species with open pangenomes present further challenges to ML-based
483 prediction of antibiotic resistance. Increased resistance mechanism complexity and
484 greater inter-isolate variation in resistance mechanisms require more intensive sampling
485 to capture a significant portion of the resistome[56]. On the technical side, even for heavily
486 sampled species, when using whole genome feature sets, the number of genetic features
487 (e.g., k-mers or SNPs) will always be much larger than the number of observations
488 (isolates), increasing the risk of overfitting[12]. This can be particularly problematic in
489 species with open pangenomes, as the ratio of genetic features to the number of genomes
490 is larger and the number of unique genetic features per number of genomes does not
491 plateau, even with heavy sampling. By comparing classifier performance in predicting CIP
492 NS across gonococci, *K. pneumoniae*, and *A. baumannii*, we show that classifiers
493 generally did not perform as well for species with open genomes (*K. pneumoniae* or *A.*
494 *baumannii*) as for gonococci. Further, while a single GyrA mutation could explain the
495 majority of CIP NS across all species evaluated here, unlike in gonococci and *A.*
496 *baumannii* where this mutation explained ≥97% of CIP NS, 14% of CIP NS in *K.*

497 *pneumoniae* could not be explained by this mutation, suggesting increased CIP
498 resistance mechanism diversity and/or complexity in this species. While increased
499 sampling, different methods, and/or finer tuning of hyperparameters may yield increased
500 prediction accuracy for drug resistance in species with open genomes (e.g., Nguyen et
501 al., 2018 reported a mean bACC of 98.5% using a decision tree-based extreme gradient
502 boosting regression model to predict CIP MICs for the *K. pneumoniae* strains assessed
503 here[14]), our results demonstrate clear variation in potential limitations of genotype-to-
504 resistance-phenotype models across different species.
505
506 Given the biological and epidemiological disparities associated with resistance to
507 different drugs in different clinical populations and bacterial species, and their evident
508 impact on performance of predictive models, successful implementation of genotype-
509 based resistance diagnostics will likely require sustained comprehensive sampling,
510 customized modeling, and incorporation of feedback mechanisms based on treatment
511 outcome data. Further evaluation of additional ML methods and datasets may reveal
512 more quantitative requirements and limitations associated with the application of
513 genotype-to-resistance-phenotype predictive modeling in the clinical setting.
514
515 **Acknowledgements**
516 ALH and YHG are supported by the Richard and Susan Smith Family Foundation
517 (http://www.smithfamilyfoundation.net/) and a National Institutes of Health R01 AI132606
518 (https://www.nih.gov/). This work was also supported by Wellcome (grant 098051 to the
519 Wellcome Sanger Institute). NEW was funded through National Institutes of Health grant
520 U01CA207167. The funders had no role in study design, data collection and analysis,

521 decision to publish, or preparation of the manuscript. Its contents are solely the
522 responsibility of the authors and do not necessarily represent the official views of the
523 National Institutes of Health. We thank Jung-Eun Shin, Mark Labrador, and members of
524 the Grad Lab for helpful discussion, and Julie Schillinger and Preeti Pathela for assistance
525 identifying, selecting, and characterizing the isolates from New York City.
526

527 **Author contributions**

528 ALH and YHG conceived of the study. ALH performed the analyses, and ALH and YHG
529 drafted the manuscript. JLR provided samples, including culturing specimens, isolating
530 DNA, and testing antibiotic susceptibility. ALH, YHG, LSB, and NW interpreted the data.
531 All authors contributed to the writing of the manuscript.
532

533 **References:**

534 1 The Review on Antimicrobial Resistance. 2016. Tackling drug-resistant infections
535 globally: final report and recommendations. London, United Kingdom.
536 2 Zumla, A. *et al.* Rapid point of care diagnostic tests for viral and bacterial
537 respiratory tract infections--needs, advances, and future prospects. *Lancet Infect*
538 *Dis* **14**, 1123-1135, doi:10.1016/S1473-3099(14)70827-8 (2014).
539 3 Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming
540 clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* **13**, 601-
541 612, doi:10.1038/nrg3226 (2012).

542 4 Walker, T. M. *et al.* Whole-genome sequencing for prediction of Mycobacterium
543 tuberculosis drug susceptibility and resistance: a retrospective cohort study.
544 *Lancet Infect Dis* **15**, 1193-1202, doi:10.1016/S1473-3099(15)00062-6 (2015).
545 5 Rigouts, L. *et al.* Rifampin resistance missed in automated liquid culture system
546 for Mycobacterium tuberculosis isolates with specific rpoB mutations. *J Clin*
547 *Microbiol* **51**, 2641-2645, doi:10.1128/JCM.02741-12 (2013).
548 6 Mason, A. *et al.* Accuracy of Different Bioinformatics Methods in Detecting
549 Antibiotic Resistance and Virulence Factors from Staphylococcus aureus Whole-
550 Genome Sequences. *J Clin Microbiol* **56**, doi:10.1128/JCM.01815-17 (2018).
551 7 Yang, Y. *et al.* Machine learning for classifying tuberculosis drug-resistance from
552 DNA sequencing data. *Bioinformatics* **34**, 1666-1671,
553 doi:10.1093/bioinformatics/btx801 (2018).
554 8 Pesesky, M. W. *et al.* Evaluation of Machine Learning and Rules-Based
555 Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative
556 Bacilli from Whole Genome Sequence Data. *Front Microbiol* **7**, 1887,
557 doi:10.3389/fmicb.2016.01887 (2016).
558 9 Li, Y. *et al.* Validation of beta-lactam minimum inhibitory concentration predictions
559 for pneumococcal isolates with newly encountered penicillin binding protein (PBP)
560 sequences. *BMC Genomics* **18**, 621, doi:10.1186/s12864-017-4017-7 (2017).
561 10 Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence
562 data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat Commun* **6**,
563 10063, doi:10.1038/ncomms10063 (2015).

564 11 Drouin, A. *et al.* Predictive computational phenotyping and biomarker discovery
565 using reference-free genome comparisons. *BMC Genomics* **17**, 754,
566 doi:10.1186/s12864-016-2889-6 (2016).
567 12 Drouin, A. *et al.* Interpretable genotype-to-phenotype classifiers with performance
568 guarantees. *bioRxiv*, doi: http://dx.doi.org/10.1101/388348 (2018).
569 13 Davis, J. J. *et al.* Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci*
570 *Rep* **6**, 27930, doi:10.1038/srep27930 (2016).
571 14 Nguyen, M. *et al.* Developing an in silico minimum inhibitory concentration panel
572 test for Klebsiella pneumoniae. *Sci Rep* **8**, 421, doi:10.1038/s41598-017-18972-w
573 (2018).
574 15 Nguyen, M. *et al.* Using machine learning to predict antimicrobial minimum
575 inhibitory concentrations and associated genomic features for nontyphoidal
576 Salmonella. *J Clin Microbiol*, doi:10.1101/380782 (2018).
577 16 Santerre, J. W., Davis, J. J., Xia, F. & Stevens, R. Machine Learning for
578 Antimicrobial Resistance. *arXiv e-prints*, doi:arXiv:1607.01224 (2016).
579 17 Moradigaravand, D. *et al.* Prediction of antibiotic resistance in Escherichia coli from
580 large-scale pan-genome data. *PLoS Comput Biol* **14**, e1006258,
581 doi:10.1371/journal.pcbi.1006258 (2018).
582 18 Gordon, N. C. *et al.* Prediction of Staphylococcus aureus antimicrobial resistance
583 by whole-genome sequencing. *J Clin Microbiol* **52**, 1182-1191,
584 doi:10.1128/JCM.03117-13 (2014).
585 19 Marchland, M. & Shawe-Taylor, J. The set covering machine. *Journal of Machine*
586 *Learning Research* **3**, 723-746 (2002).

587 20 Breiman, L. Random forests. *Machine Learning* **45**, 5-32 (2001).

588 21 Hemarajata, P., Yang, S., Soge, O. O., Humphries, R. M. & Klausner, J. D.
589 Performance and Verification of a Real-Time PCR Assay Targeting the gyrA Gene
590 for Prediction of Ciprofloxacin Resistance in Neisseria gonorrhoeae. *J Clin*
591 *Microbiol* **54**, 805-808, doi:10.1128/JCM.03032-15 (2016).

592 22 Siedner, M. J. *et al.* Real-time PCR assay for detection of quinolone-resistant
593 Neisseria gonorrhoeae in urine samples. *J Clin Microbiol* **45**, 1250-1254,
594 doi:10.1128/JCM.01909-06 (2007).

595 23 Grad, Y. H. *et al.* Genomic Epidemiology of Gonococcal Resistance to Extended-
596 Spectrum Cephalosporins, Macrolides, and Fluoroquinolones in the United States,
597 2000-2013. *J Infect Dis* **214**, 1579-1587, doi:10.1093/infdis/jiw420 (2016).

598 24 Wadsworth, C. B., Arnold, B. J., Sater, M. R. A. & Grad, Y. H. Azithromycin
599 Resistance through Interspecific Acquisition of an Epistasis-Dependent Efflux
600 Pump Component and Transcriptional Regulator in Neisseria gonorrhoeae. *MBio*
601 **9**, doi:10.1128/mBio.01419-18 (2018).

602 25 Eyre, D. W. *et al.* WGS to predict antibiotic MICs for Neisseria gonorrhoeae. *J*
603 *Antimicrob Chemother* **72**, 1937-1947, doi:10.1093/jac/dkx067 (2017).

604 26 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its
605 applications to single-cell sequencing. *J Comput Biol* **19**, 455-477,
606 doi:10.1089/cmb.2012.0021 (2012).

607 27 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment
608 tool for genome assemblies. *Bioinformatics* **29**, 1072-1075,
609 doi:10.1093/bioinformatics/btt086 (2013).

610 28 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
611 alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-
612 2836(05)80360-2 (1990).
613 29 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
614 throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
615 30 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
616 MEM. *arXiv e-prints*, doi:arXiv:1303.3997 (2013).
617 31 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant
618 detection and genome assembly improvement. *PLoS One* **9**, e112963,
619 doi:10.1371/journal.pone.0112963 (2014).
620 32 Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for
621 High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**, 1-17,
622 doi:doi:10.18637/jss.v077.i01 (2017).
623 33 Rizk, G., Lavenier, D. & Chikhi, R. DSK: k-mer counting with very low memory
624 usage. *Bioinformatics* **29**, 652-653, doi:10.1093/bioinformatics/btt020 (2013).
625 34 Earle, S. G. *et al.* Identifying lineage effects when controlling for population
626 structure improves power in bacterial association studies. *Nat Microbiol* **1**, 16041,
627 doi:10.1038/nmicrobiol.2016.41 (2016).
628 35 Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic
629 basis of bacterial phenotypes. *Nat Commun* **7**, 12797, doi:10.1038/ncomms12797
630 (2016).

631 36 The European Committee on Antimicrobial Susceptibility Testing. *Breakpoint*
632 *tables for interpretation of MICs and zone diameters. Version 8.1, 2018.* URL:
633 http://www.eucast.org/.
634 37 Clinical and Laboratory Standards Institute. CLSI M100: Performance Standards
635 for Antimicrobial Susceptibility Testing, 29th Edition (2019).
636 38 Harris, S. R. *et al.* Public health surveillance of multidrug-resistant clones of
637 Neisseria gonorrhoeae in Europe: a genomic survey. *Lancet Infect Dis* **18**, 758-
638 768, doi:10.1016/S1473-3099(18)30225-1 (2018).
639 39 Yahara, K. *et al.* Genomic surveillance of Neisseria gonorrhoeae to investigate the
640 distribution and evolution of antimicrobial-resistance determinants and lineages.
641 *Microb Genom* **4**, doi:10.1099/mgen.0.000205 (2018).
642 40 Demczuk, W. *et al.* Genomic Epidemiology and Molecular Resistance
643 Mechanisms of Azithromycin-Resistant Neisseria gonorrhoeae in Canada from
644 1997 to 2014. *J Clin Microbiol* **54**, 1304-1313, doi:10.1128/JCM.03195-15 (2016).
645 41 Niehaus, K. E., Walker, T. M., Crook, D. W. & Clifton, T. E. A. P. A. in *IEEE-EMBS*
646 *International Conference on Biomedical and Health Informatics (BHI)* 618-621
647 (2014).
648 42 Olesen, S. W. *et al.* Azithromycin susceptibility in Neisseria gonorrhoeae and
649 seasonal macrolide use. *J Infect Dis* **jiy551** (2018).
650 43 Unemo, M. & Shafer, W. M. Antibiotic resistance in Neisseria gonorrhoeae: origin,
651 evolution, and lessons learned for the future. *Ann N Y Acad Sci* **1230**, E19-28,
652 doi:10.1111/j.1749-6632.2011.06215.x (2011).

653 44 Andre, E. *et al.* Novel rapid PCR for the detection of Ile491Phe rpoB mutation of
654 Mycobacterium tuberculosis, a rifampicin-resistance-conferring mutation
655 undetected by commercial assays. *Clin Microbiol Infect* **23**, 267 e265-267 e267,
656 doi:10.1016/j.cmi.2016.12.009 (2017).
657 45 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for
658 association studies. *Nat Genet* **44**, 821-824, doi:10.1038/ng.2310 (2012).
659 46 Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive
660 selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* **45**, 1183-1189,
661 doi:10.1038/ng.2747 (2013).
662 47 Prideaux, B. *et al.* The association between sterilizing activity and drug distribution
663 into tuberculosis lesions. *Nat Med* **21**, 1223-1227, doi:10.1038/nm.3937 (2015).
664 48 Tamma, P. D. *et al.* Outcomes of children with enterobacteriaceae bacteremia with
665 reduced susceptibility to ceftriaxone: do the revised breakpoints translate to
666 improved patient outcomes? *Pediatr Infect Dis J* **32**, 965-969,
667 doi:10.1097/INF.0b013e31829043b3 (2013).
668 49 Bhat, S. V. *et al.* Failure of current cefepime breakpoints to predict clinical
669 outcomes of bacteremia caused by gram-negative organisms. *Antimicrob Agents*
670 *Chemother* **51**, 4390-4395, doi:10.1128/AAC.01487-06 (2007).
671 50 Tam, V. H. *et al.* Outcomes of bacteremia due to Pseudomonas aeruginosa with
672 reduced susceptibility to piperacillin-tazobactam: implications on the
673 appropriateness of the resistance breakpoint. *Clin Infect Dis* **46**, 862-867,
674 doi:10.1086/528712 (2008).

675 51 Colangeli, R. *et al.* Bacterial Factors That Predict Relapse after Tuberculosis
676 Therapy. *N Engl J Med* **379**, 823-833, doi:10.1056/NEJMoa1715849 (2018).
677 52 Humphries, R. M. *et al.* CLSI Methods Development and Standardization Working
678 Group Best Practices for Evaluation of Antimicrobial Susceptibility Tests. *J Clin*
679 *Microbiol* **56**, doi:10.1128/JCM.01934-17 (2018).
680 53 Chen, M. L. *et al.* Deep Learning Predicts Tuberculosis Drug Resistance Status
681 from Whole-Genome Sequencing Data. *bioRxiv*,
682 doi:https://doi.org/10.1101/275628 (2018).
683 54 Jeni, L. A., Cohn, J. F. & De La Torre, F. in *2013 Humaine Association Conference*
684 *on Affective Computing and Intelligent Interaction* 245-251 (2013).
685 55 FDA. Class II Special Controls Guidance Document: Antimicrobial Susceptibility
686 Test (AST) Systems; Guidance for Industry and FDA (Food and Drug
687 Administration, Rockville, MD, 2009).
688 56 Jeukens, J. *et al.* Genomics of antibiotic-resistance prediction in Pseudomonas
689 aeruginosa. *Ann N Y Acad Sci*, doi:10.1111/nyas.13358 (2017).
690 57 De Silva, D. *et al.* Whole-genome sequencing to determine transmission of
691 Neisseria gonorrhoeae: an observational study. *Lancet Infect Dis* **16**, 1295-1303,
692 doi:10.1016/S1473-3099(16)30157-8 (2016).
693 58 Demczuk, W. *et al.* Whole-genome phylogenomic heterogeneity of Neisseria
694 gonorrhoeae isolates with decreased cephalosporin susceptibility collected in
695 Canada between 1989 and 2013. *J Clin Microbiol* **53**, 191-200,
696 doi:10.1128/JCM.02589-14 (2015).

697 59 Lee, R. S. *et al.* Genomic epidemiology and antimicrobial resistance of Neisseria
698 gonorrhoeae in New Zealand. *J Antimicrob Chemother* **73**, 353-364,
699 doi:10.1093/jac/dkx405 (2018).
700 60 Lesho, E. P. *et al.* The antimicrobial resistance monitoring and research (ARMoR)
701 program: the US Department of Defense response to escalating antimicrobial
702 resistance. *Clin Infect Dis* **59**, 390-397, doi:10.1093/cid/ciu319 (2014).

703

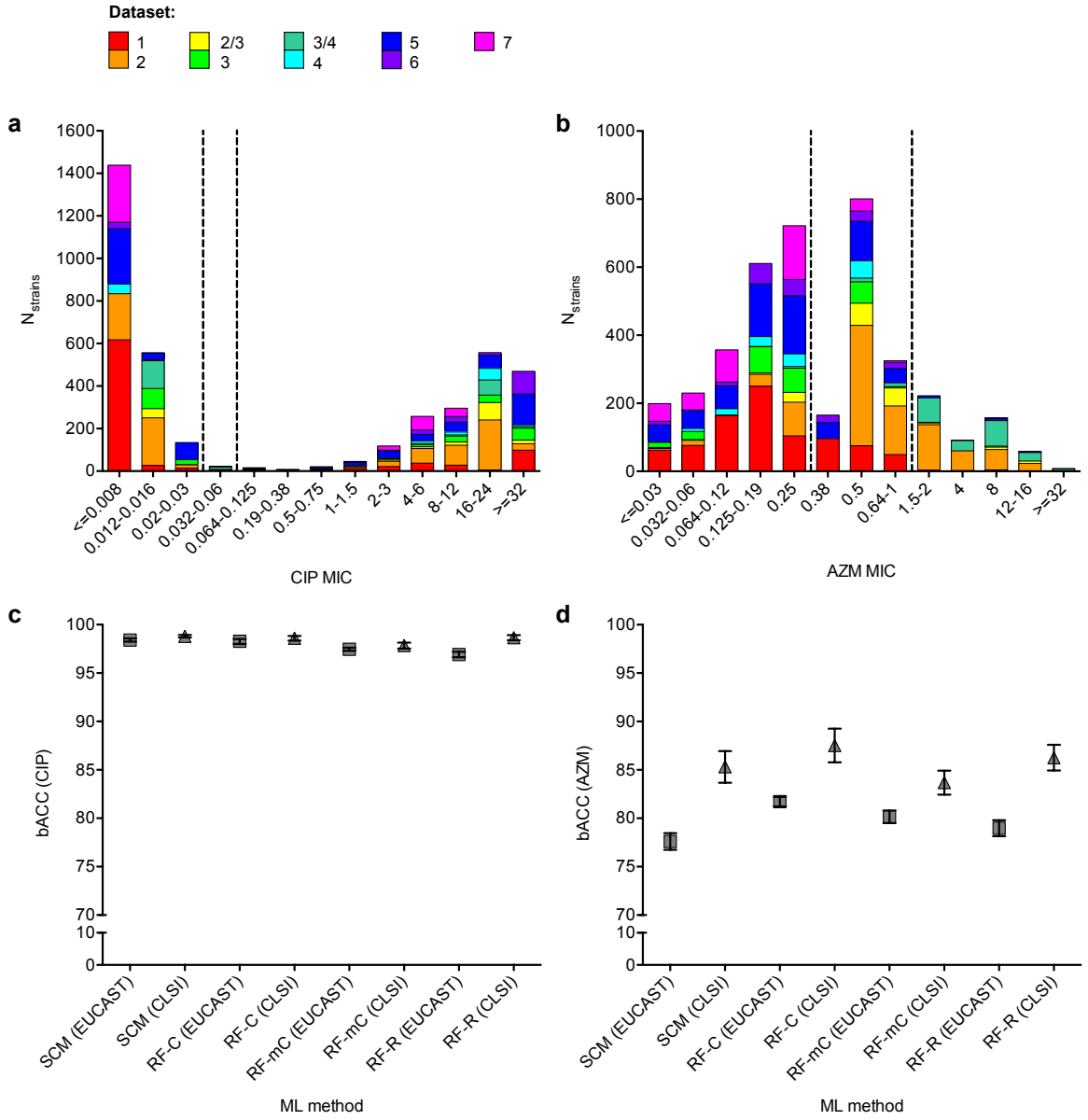
704
705 **Tables and Figures:**
706
707 **Table 1.** Summary of datasets.

Species	Dataset	SRA Study ID/Reference	$N_{samples}$	Temporal range	Geographic range	Sampling approach
N. gonorrhoeae	1	ERP011192	886	2011-2015	New York, NY (US)	Survey from citywide clinics
	2	ERP008891, ERP001405, ERP000144 [23]	1102	2000-2013	National (US)	Survey from nationwide clinics; male patients only; enriched for CFX resistance
	3	SRP065041, ERP008891, SRP072971 [25]	671	2004-2014	International (UK, Canada, US)	Surveys from Brighton, UK [57] and nationwide sites in Canada [40,58] and the US [23]; Canadian samples enriched for CRO and AZM resistance; US samples enriched for CFX resistance; US samples from male patients only
	4	SRP050190, SRP065041 [40,58]	383	1989-2014	National (Canada)	Surveys from nationwide sites in Canada; enriched for CRO and AZM resistance
	5	ERP010312 [38]	714	2013	International (Europe)	Survey from clinics and hospitals across 21 European countries
	6	DRP004052 [39]	204	2015	National (Japan)	Survey from clinics in Kyoto and Osaka; male patients only

	7	SRP111927[59]	398	2014-2015	National (New Zealand)	Survey from nationwide diagnostic labs
K. pneumoniae	8	SRP102664, SRP110988, SRP116139[14]	1560	2011-2017	Houston, TX (US)	Survey from citywide hospital system; enriched for β-lactam resistance
A. baumannii	9	SRP065910[60]	702	2000-2012	National (US)	Survey from clinics and hospitals within the US military healthcare system

708 CFX, cefixime; CRO, ceftriaxone; AZM, azithromycin
709

710



711

712 **Figure 1. Differential performance of machine learning-based prediction models for**

713 **ciprofloxacin and azithromycin resistance in gonococci**. Histograms showing the

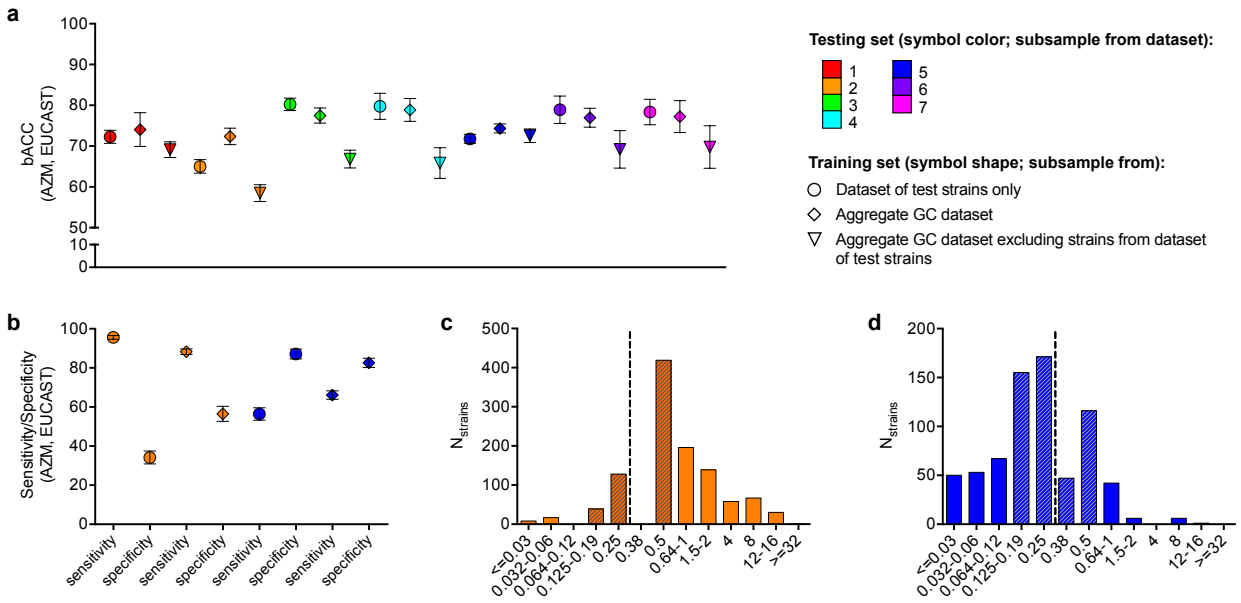
714 distributions of **(a)** ciprofloxacin (CIP) and **(b)** azithromycin (AZM) MICs in the gonococcal

715 isolates assessed here. Bar color indicates the study or studies associated with the

716 isolates. Dashed lines indicate the **(a)** EUCAST and CLSI breakpoints for non-

717 susceptibility (NS, >0.03 μg/mL and >0.06 μg/mL, respectively) for CIP and the **(b)**

718 EUCAST and CLSI breakpoints for non-susceptibility (>0.25 μg/mL and >1 μg/mL,
719 respectively) for AZM. Mean balanced accuracy (bACC) with 95% confidence intervals of
720 predictive models for **(c)** CIP NS and **(d)** AZM NS trained and tested on the aggregate
721 gonococcal dataset. SCM, set covering machine; RF-C, random forest classification; RF-
722 mC, random forest multi-class classification; RF-R, random forest regression.



723

724 **Figure 2. Differential performance of random forest classifiers trained on different**

725 **datasets. (a)** Mean balanced accuracy (bACC) with 95% confidence intervals of

726 predictive models for gonococci (GC) azithromycin (AZM) non-susceptibility based on the

727 EUCAST breakpoint. **(b)** Mean sensitivity and specificity with 95% confidence intervals of

728 predictive models for GC AZM non-susceptibility in datasets 2 and 5. Histograms showing

729 the distributions of AZM MICs in **(c)** dataset 2 and **(d)** dataset 5. Symbol colors in **(a)** and

730 **(b)** indicate the dataset from which the testing set was derived, while symbol shape in **(a)**

731 and **(b)** indicates the dataset from which the training set was derived. Hatching in **(c)** and

732 **(d)** indicates MICs within one doubling dilution of the EUCAST breakpoint (designated by

733 dashed lines).

734

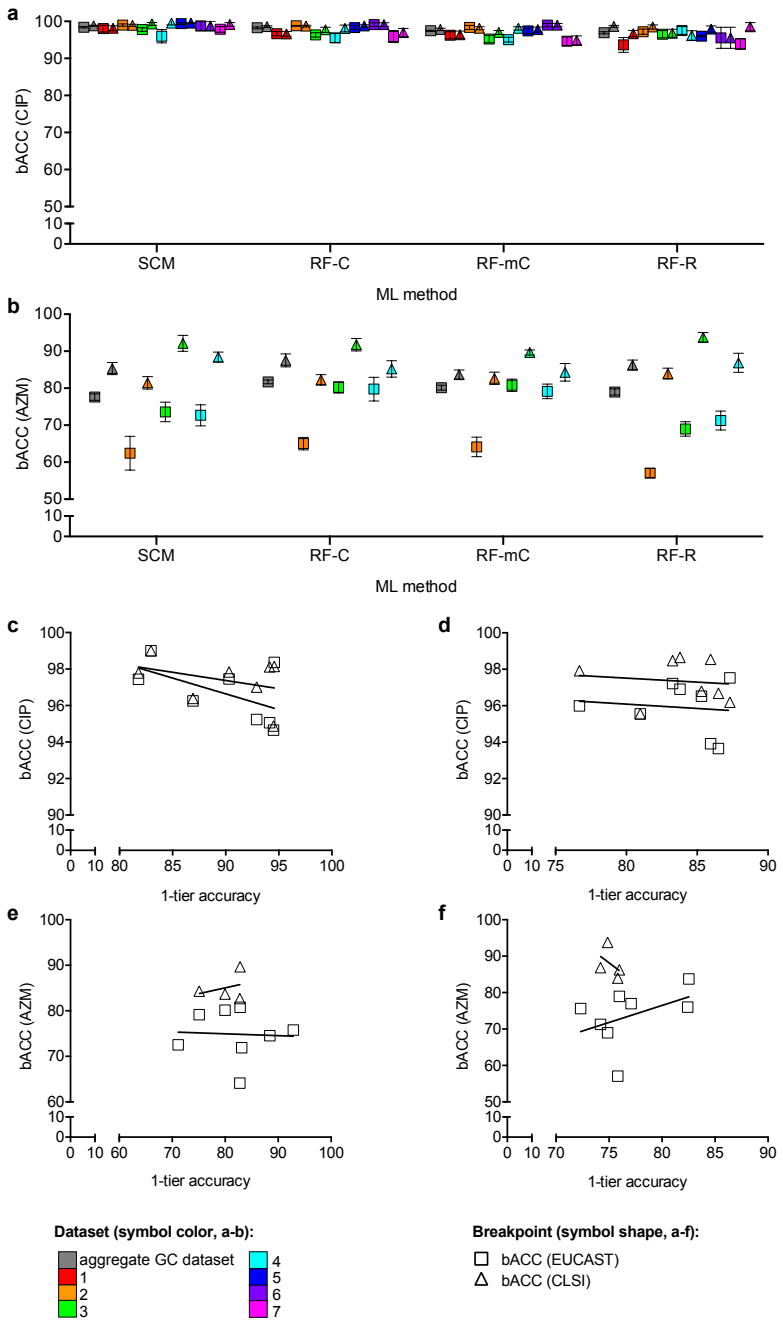


Figure 3. Differential performance of machine learning-based prediction models based on different resistance metrics in gonococci. Mean balanced accuracy (bACC) with 95% confidence intervals of predictive models for **(a)** ciprofloxacin non-susceptibility (CIP NS) across all datasets and **(b)** azithromycin (AZM) NS for all datasets for which both NS breakpoints were evaluated. Scatter plots comparing the mean 1-tier accuracy

741 to the mean bACC for each gonococcal dataset derived from **(c-d)** CIP and **(e-f)** AZM
742 MIC prediction models by **(c,e)** random forest multi-class classification and **d,f** random
743 forest regression. Symbol colors in **(a-b)** indicate the datasets from which the training and
744 testing sets were derived. Symbol shapes in **(a-f)** indicate the NS breakpoint. The line of
745 best fit for each of the breakpoints is indicated in **(c-f)**. SCM, set covering machine; RF-
746 C, random forest binary classification; RF-mC, random forest multi-class classification;
747 RF-R, random forest regression.

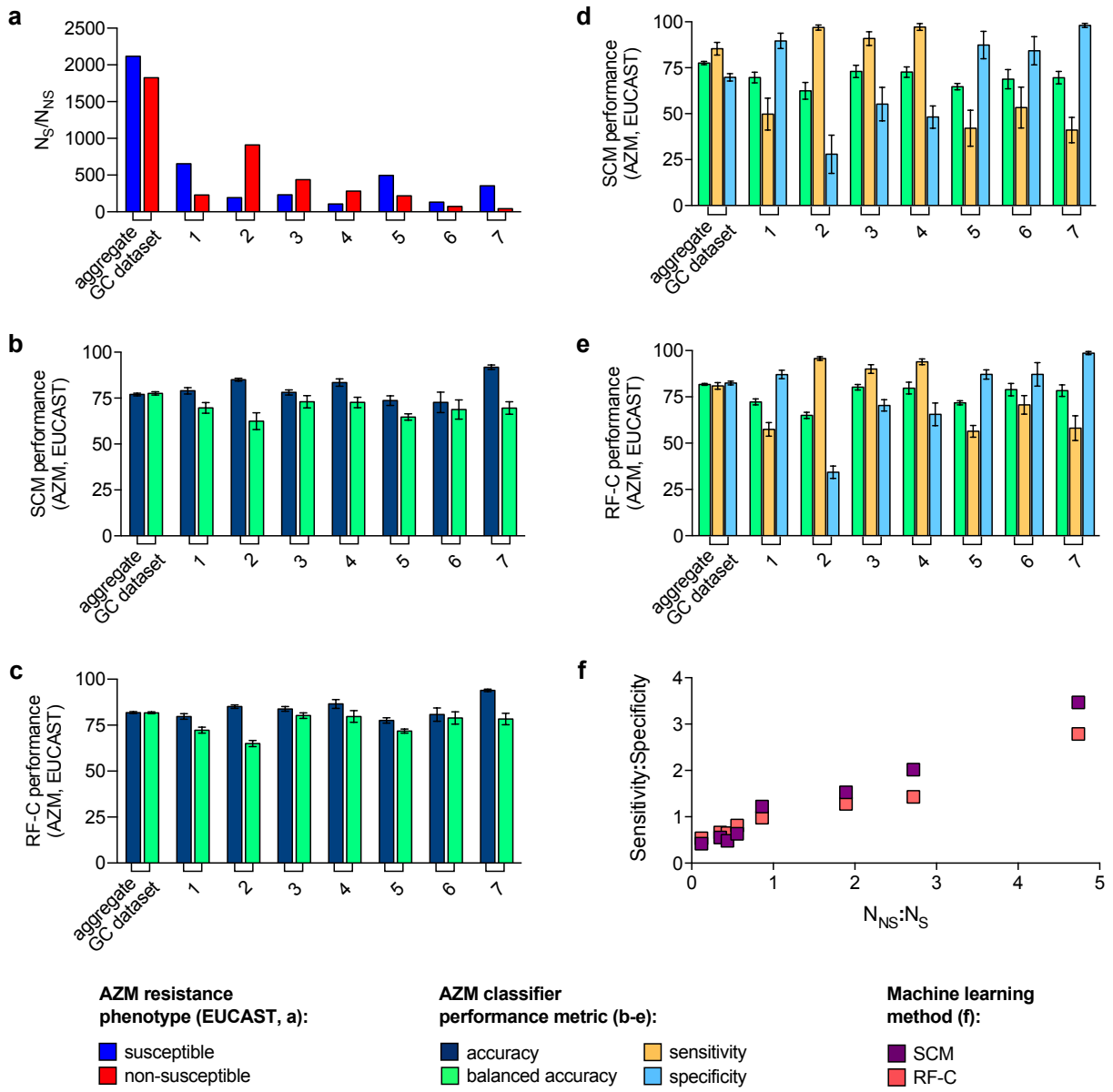


Figure 4. Differential performance of predictive models of resistance across different performance metrics. (a) Distribution of azithromycin (AZM) susceptible (S) and non-susceptible (NS) strains by the EUCAST breakpoint in each gonococcal dataset. Mean accuracy and balanced accuracy (bACC) with 95% confidence intervals achieved by **(b)** set covering machine (SCM) and **(c)** random forest classification (RF-C) models for AZM NS by the EUCAST breakpoint across gonococcal datasets. Mean bACC, sensitivity, and specificity with 95% confidence intervals achieved by **(d)** SCM and **(e)** RF-

756 C models for AZM NS by the EUCAST breakpoint across gonococcal datasets. **(f)** Scatter
757 plot showing the relationship between the ratio of NS strains to S strains in each dataset
758 and the ratio of sensitivity to specificity achieved by SCM and RF-C methods.
759

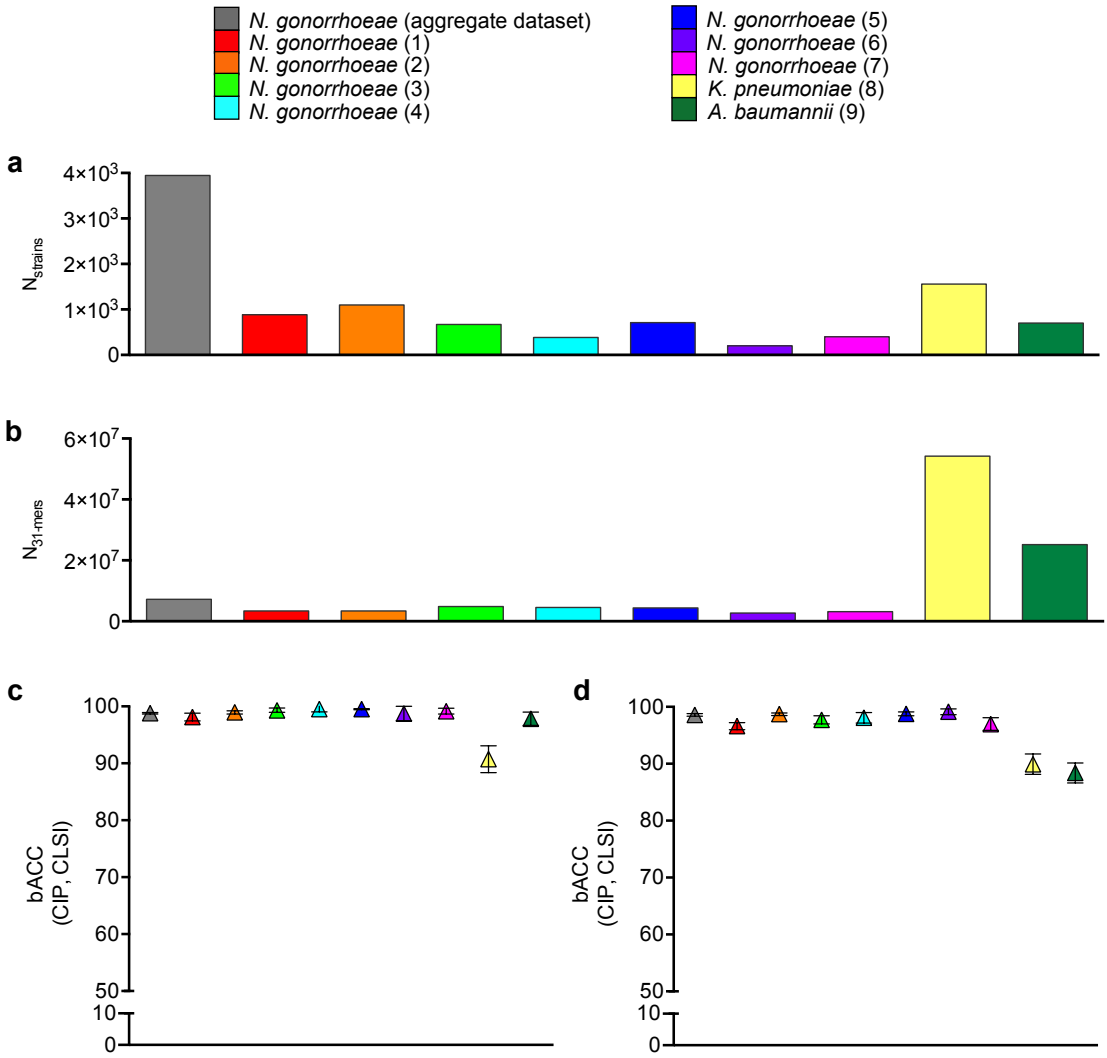


Figure 5. *K. pneumoniae* and *A. baumannii* datasets are associated with higher genetic diversity and lower performance of resistance prediction models. Number of **a)** strains and **b)** unique 31-mers in each dataset. Mean balanced accuracy (bACC) with 95% confidence intervals achieved by **c)** set covering machine and **d)** random forest classification models for ciprofloxacin (CIP) NS by the CLSI breakpoints across gonococci, *K. pneumoniae,* and *A. baumannii* datasets.