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[2,3], Jennifer L. Rakeman[4], Simon R. Harris[5], Yonatan H. Grad[1,6,*]

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 Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, 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 Microbiotica Ltd, Biodata Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

6 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 data 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, 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.

Author Summary:

Machine learning-based prediction of antibiotic resistance from bacterial genome sequences represents a promising tool to rapidly determine the antibiotic susceptibility profile of clinical isolates and reduce the morbidity and mortality resulting from inappropriate and ineffective treatment. However, while there has been much focus on demonstrating the diagnostic potential of these modeling approaches, there has been little assessment of potential caveats and prerequisites associated with implementing predictive models of drug resistance in the clinical setting. Our results highlight significant biological and technical challenges facing the application of machine learning-based prediction of antibiotic resistance as a diagnostic tool. By outlining specific factors affecting model performance, our findings provide a framework for future work on modeling drug resistance and underscore the necessity of continued comprehensive sampling and reporting of treatment outcome data for building reliable and sustainable diagnostics.

59 **Introduction:**

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

70 Molecular assays offer improved speed compared to gold-standard phenotypic
71 tests and are of particular interest because of their promise of high accuracy for the
72 prediction of AMR phenotype based on genotype (2, 3). Approaches for predicting
73 resistance phenotypes from genetic features include direct association (*i.e.*, using the
74 presence or absence of genetic variants known to be associated with resistance to infer
75 a resistance phenotype) and the application of predictive models derived from machine
76 learning (ML) algorithms. Direct association approaches can offer simple, inexpensive,
77 and often highly accurate resistance assays for some drugs/species (2) and may even
78 provide more reliable predictions of resistance phenotype than phenotypic testing (4-6).
79 However, these approaches are limited by the availability of well-curated and up-to-date
80 panels of resistance variants, as well as the diversity and complexity of resistance
81 mechanisms. ML strategies can facilitate modeling of more complex, diverse, and/or

82 under-characterized resistance mechanisms, thus outperforming direct association for
83 many drugs/species (7-9). With the increasing speed and decreasing cost of sequencing
84 and computation, ML approaches can be applied to genome-wide feature sets (8, 10-18),
85 ideally obviating the need for comprehensive *a priori* knowledge of resistance loci.

86 While prediction of antibiotic resistance phenotypes from ML-derived models
87 based on genomic features has become increasingly prominent as a promising diagnostic
88 tool (8, 11-15, 17), there has been no systematic evaluation of factors that may influence
89 performance of such models and their implications in the clinical setting. The extent to
90 which ML model accuracy varies by antibiotic is unclear, as is the impact of sampling bias
91 on model performance. It is further unclear what the most relevant resistance metric (*i.e.*,
92 minimum inhibitory concentration [MIC] or categorical report of susceptibility) for such a
93 diagnostic might be and how amenable different species might be to genotype-to-
94 phenotype modeling of antibiotic resistance.

95 We used set covering machine (SCM) (19) and random forest (RF) (20)
96 classification as well as RF regression algorithms to build and test predictive models with
97 seven gonococcal datasets for which whole genome sequences (WGS) and ciprofloxacin
98 (CIP) and azithromycin (AZM) MICs were available. AZM is currently part of the
99 recommended treatment regimen for gonococcal infections, and with the development of
100 resistance diagnostics, CIP may represent a viable treatment option (21-23). While the
101 majority of CIP resistance in gonococci can be attributed to *gyrA* mutations, AZM
102 resistance is associated with more diverse and complex resistance mechanisms (23, 24),
103 offering an opportunity to evaluate ML methods across drugs with distinct pathways to
104 resistance. The range of datasets and sampling frames enables assessment of sampling

105 bias on model reliability. Further, the availability of MICs, as well as distinct European
106 Committee on Antibiotic Susceptibility Testing (EUCAST) and Clinical and Laboratory
107 Standards Institute (CLSI) breakpoints, for these drugs allows for evaluation of predictive
108 models based on different resistance metrics. Finally, extension of these analyses to
109 *Klebsiella pneumoniae* and *Acinetobacter baumannii* datasets for which WGS and CIP
110 MICs were available allows for assessment of model performance for the same drug in
111 species with open pangenomes (25, 26), which may be more difficult to model given the
112 increased genomic diversity and potential resistance mechanism diversity and complexity
113 (47).

114 Our results demonstrate that using ML to predict antibiotic resistance phenotypes
115 from WGS data yields variable results across drugs, datasets, resistance metrics, and
116 species. While more comprehensive assessment of different methods will be required to
117 build the most accurate and reliable models, we suggest that tailored modeling for
118 individual drugs, species, and clinical populations may be necessary to successfully
119 leverage these ML-based approaches as diagnostic tools. We further suggest that
120 continuing surveillance, isolate collection, and reporting of WGS, MIC phenotypes, and
121 treatment outcomes will be crucial to the sustainability of any such molecular diagnostics.
122

123 **Results:**

124 **Accuracy of ML-based prediction of resistance phenotypes varies by antibiotic.**

125 Given the distinct MIC distributions and distinct pathways to resistance for CIP and AZM
126 in gonococci, these two drugs enable evaluation of drug-specific performance of ML-
127 based resistance prediction models. CIP MICs in surveys of clinical gonococcal isolates

128 are bimodally distributed, with the majority of isolates having MICs well above or below
129 the non-susceptibility (NS) breakpoints, while the majority of reported AZM MICs in
130 gonococci are closer to the NS breakpoints (https://mic.eucast.org/Eucast2). These
131 trends were recapitulated in the gonococcal isolates assessed here (**Fig 1a-b**). Further,
132 the vast majority of CIP resistance in gonococci observed to date is explained by
133 mutations in *gyrA* and *parC* and has spread predominantly through clonal expansion,
134 generally resulting in MICs ≥ 1 μg/mL (23, 27). In contrast, AZM resistance in gonococci
135 has arisen many times *de novo* through multiple pathways, many of which remain under-
136 characterized and are associated with lower-level resistance (23, 27, 28). As expected,
137 the GyrA S91F mutation alone predicts NS to CIP by both EUCAST and CLSI breakpoints
138 in the aggregate gonococcal dataset assessed here with ≥98% sensitivity and ≥99%
139 specificity (**Table S1**). AZM NS showed lower values for these metrics, indicating it was
140 not as well explained by known resistance variants, with extensive contributions from
141 uncharacterized mechanisms and/or multifactorial interactions (**Table S2**).

142 We next trained and evaluated ML-based predictive models for CIP and AZM
143 resistance in gonococci (**Table S3**). By all ML methods and breakpoints, CIP NS was
144 predicted with significantly higher balanced accuracy (bACC) than AZM NS in the
145 aggregate gonococcal dataset ($P < 0.0001$, **Fig 1c-d, Tables S4-S5**): CIP NS was
146 predicted with mean bACC ≥93% across all methods, breakpoints, and datasets, whereas
147 mean bACC for AZM NS classification ranged from 57% to 94% (**Tables S4-S5**).
148 Variation in model performance across antibiotics has been attributed to different
149 proportions of susceptible (S) and NS isolates (7, 14, 15); however, by the EUCAST
150 breakpoints, the aggregate gonococcal dataset as well as some of the individual datasets

151 had nearly identical proportions of CIP and AZM susceptible and non-susceptible isolates,
152 demonstrating that variable representation of S and NS isolates alone cannot explain
153 reduced performance of AZM models compared to CIP.

154 We tested whether the poorer performance for AZM may be attributable to the
155 large fraction of isolates with MICs around the breakpoint. Removing strains with AZM
156 MICs that were ≤2 doubling dilutions of the NS breakpoints from the aggregate
157 gonococcal dataset (**Table S6**) yielded AZM MIC distributions similar to those of CIP (**Fig**
158 **S1a-b**). Analysis of this restricted dataset resulted in higher performance of SCM and RF
159 AZM NS classifiers compared to those trained and tested on the full aggregate
160 gonococcal dataset (**Fig S1c**). However, bACC of AZM classifiers trained and tested on
161 the restricted datasets was still significantly lower than bACC of the CIP NS classifiers (P
162 < 0.0001 and $P < 0.003$ for classifiers based on the EUCAST and CLSI breakpoints,
163 respectively), suggesting that both MIC distribution and additional drug-specific factors
164 can influence performance of resistance classifiers.

165

166 **Sampling bias in training and testing data skews resistance model performance.**
167 The diversity of resistance mechanisms for AZM in gonococci offers an opportunity to
168 evaluate the effects of sampling bias on model performance. The sampling frames for the
169 seven gonococcal datasets ranged geographically from citywide to international and
170 temporally from a single year to >20 years, and several datasets were enriched for AZM
171 resistance (11, 29) (**Table 1**). The distributions of both AZM MICs and known resistance
172 mechanisms across datasets (**Fig 1b, Table S2**) and the variable performance of AZM
173 resistance models across datasets (**Table S5**) suggest that AZM resistance mechanisms

174 are differentially distributed across the sampled clinical populations. Further, the higher
175 performance of many SCM and RF-based AZM classifiers on training data compared to
176 test sets (**Table S5**) suggests that potentially due to a lack of signal, AZM models are
177 incorporating substantial noise or confounding factors, which may be population-specific.
178 To assess the impact of sampling on model reliability, the performance of RF classifiers
179 in prediction of AZM NS phenotypes were compared across multiple training and testing
180 sets. These include classifiers trained on subsamples of isolates from a single dataset,
181 classifiers trained on the aggregate gonococcal dataset, and classifiers trained on the
182 aggregate gonococcal dataset excluding isolates from the same dataset as the testing
183 set (**Table S6**). Given the low representation of AZM NS strains by the CLSI breakpoint
184 in many datasets, these analyses were only performed using the EUCAST breakpoint.

185 While it may be assumed that increased availability of paired genomic and
186 phenotypic resistance data from a broader range of clinical populations will facilitate more
187 accurate and reliable modeling (30), our results demonstrate that in predicting AZM
188 resistance phenotypes for isolates from most datasets (with the exception of datasets 2
189 and 5), performance of classifiers trained on the aggregate dataset was not significantly
190 better than performance of classifiers trained only on isolates from the dataset from which
191 the test isolates were derived ($P < 0.0001$ and $P = 0.002$ for datasets 2 and 5,
192 respectively, $P = 0.008$ for dataset 3, where the classifiers trained on the aggregate
193 dataset had lower bACC than classifiers trained only on isolates from dataset 3, and $P >$
194 0.234 for all other datasets, **Fig 2a**). Further, there was substantial variation in
195 performance of models trained on the aggregate dataset across testing sets, with models
196 achieving significantly higher bACC for strains from datasets 3 and 4 than for strains from

197 dataset 2 ($P < 0.0009$, **Fig 2a**), perhaps reflecting enrichment for AZM NS in these former
198 datasets (**Table 1**). Additionally, with the exception of dataset 5, performance of AZM
199 resistance classifiers trained only on isolates from the dataset from which the test isolates
200 were derived was significantly higher than performance of classifiers trained on the
201 aggregate dataset excluding isolates from the test dataset ($P = 0.537$ for dataset 5, $P <$
202 0.0005 for all other datasets, **Fig 2a**).

203 Performance of RF classifiers trained and tested on dataset 2 was limited by low
204 specificity, which was improved in models trained on the aggregate dataset (**Fig 2b**). The
205 low specificity achieved by RF classifiers trained and tested on this dataset is likely due
206 to the low representation of S strains, most of which were within one doubling dilution of
207 the NS breakpoint (**Fig 2c**), and thus the more comprehensive representation of negative
208 (S) data in the aggregate training set was associated with improved specificity.
209 Conversely, performance of RF classifiers trained and tested on dataset 5 was more
210 limited by low sensitivity, which was improved in models trained on the aggregate dataset
211 (**Fig 2b**). This dataset had a low representation of strains with high AZM MICs (**Fig 2d**),
212 and thus the more comprehensive representation of positive (NS) data in the aggregate
213 training set was associated with improved sensitivity in predicting AZM NS for these
214 strains. For both SCM and RF-C AZM resistance models across all datasets, there was
215 a significant positive correlation between the ratio of model sensitivity to model specificity
216 and the ratio of NS to S strains in the dataset (Pearson $r > 0.98$, $P < 0.0001$ [Pearson
217 correlation] for both SCM and RF-C, **Fig S2a**).

218 On the other hand, while representation of strains with higher AZM MICs was also
219 observed in other datasets (*i.e.*, datasets 1, 6, and 7) and was similarly reflected in the

220 sensitivity-limited performance of RF classifiers trained and tested on these datasets
221 (**Table S5**), AZM NS prediction accuracy for strains from these datasets was not improved
222 by training classifiers on the aggregate dataset. Further, even after down-sampling two
223 of the datasets with the most disparate MIC distributions, sample sizes, and model
224 performance (datasets 2 and 4) such that the number of strains and AZM MIC
225 distributions were identical between the two datasets (**Fig S2b**), there was still a
226 significant difference in AZM NS prediction accuracy of models trained and tested on
227 these different datasets (**Fig S2c,** $P < 0.004$). Together, these results demonstrate that
228 resistance model performance may be strongly associated with the distributions of both
229 resistance phenotypes and genetic features and thus can be highly population-specific.
230

231 **ML prediction models of antibiotic susceptibility / non-susceptibility outperform**
232 **MIC models**

233 Gonococcal CIP and AZM MICs were dichotomized by both EUCAST and CLSI
234 breakpoints to assess the impact of variation in MIC breakpoints on model performance.
235 As the EUCAST and CLSI breakpoints for CIP in gonococci are within a single doubling
236 dilution and the vast majority of isolates have much lower or higher CIP MICs (**Fig 1a**),
237 >99% of isolates in the aggregate dataset were consistently S or NS by both breakpoints.
238 Of the 23 isolates with MICs between the two breakpoints, 18 had MICs derived from
239 Etests of 0.032 μg/mL or 0.047 μg/mL, making their classification relative to the EUCAST
240 breakpoint of 0.03 μg/mL ambiguous. In contrast, the EUCAST and CLSI breakpoints for
241 AZM in gonococci are separated by two doubling dilutions, and for many isolates, the
242 AZM MIC was within this range (**Fig 1b**). As such, only 67% of isolates in the aggregate

243 dataset were consistently S or NS by both breakpoints. CIP NS classifier performance
244 was either identical or nearly identical for both breakpoints in the aggregate and most
245 individual gonococcal datasets (**Fig 3a**). In contrast, the bACC of AZM NS prediction by
246 both SCM and RF classifiers based on the CLSI breakpoint was significantly higher than
247 for those based on the EUCAST breakpoint across all gonococcal datasets assessed by
248 both breakpoints ($P < 0.0001$, **Fig 3b**).

249 To assess the performance of MIC prediction models relative to binary S/NS
250 resistance phenotype classifiers, RF-mC and RF-R models were trained and evaluated
251 for CIP and AZM MIC prediction in gonococci. Average exact match rates between
252 predicted and phenotypic MICs ranged from 64-86% and 54-78% by RF-mC and RF-R,
253 respectively, for CIP, and from 24-60% and 45-65%, respectively, for AZM (**Tables S4-**
254 **S5**). Average 1-tier accuracies (the percentage of isolates with predicted MICs within one
255 doubling dilution of phenotypic MICs) were substantially higher but also varied widely
256 across datasets and between the two MIC prediction methods (ranging from 82%-96%
257 and 76-87% by RF-mC and RF-R, respectively, for CIP, and from 73-94% and 73-83%,
258 respectively, for AZM; **Tables S4-S5**). There was no consistent or significant relationship
259 across the different datasets between MIC prediction accuracy (exact match or 1-tier
260 accuracy) and bACC for either drug by either MIC prediction method (**Fig 3c-f**). Further,
261 for both drugs by both breakpoints in the aggregate gonococcal dataset, binary RF-C
262 models had equivalent or significantly higher bACC than RF-mC and RF-R MIC prediction
263 models ($P > 0.175$ for AZM NS by the CLSI breakpoint by RF-C compared to RF-mC or
264 RF-R, $P < 0.017$ for all others, **Tables S4-S5**).
265

266 **Species with high genomic diversity pose challenges to ML-based antibiotic**
267 **resistance prediction**

268 Increasing genomic diversity, or an increasing ratio of genomic features (*e.g.*, *k*-mers) to
269 observations (*e.g.*, genomes), may present an additional challenge for ML-based
270 prediction of antibiotic resistance (12). To investigate ML-based antibiotic resistance
271 prediction across species with different levels of genomic diversity, SCM and RF-C were
272 used to model CIP NS in *K. pneumoniae* and *A. baumannii*, two species with genomic
273 diversity (*i.e.*, ratio of unique 31-mers to number of genomes) several times that of
274 gonococci (**Fig 4a-b**). SCM classifiers trained on and used to predict CIP NS for *K.*
275 *pneumoniae* achieved significantly lower accuracy than all of the gonococcal datasets (P
276 < 0.0001, **Fig 4c**), while SCM classifiers trained on and used to predict CIP NS for *A.*
277 *baumannii* achieved significantly lower accuracy than gonococcal datasets 3-5 and 7 (P
278 < 0.033) and roughly equivalent accuracy to gonococcal datasets 1-2 and 6, as well as
279 the aggregate gonococcal dataset ($P > 0.059$, **Fig 4c**). The performance of RF-C models
280 was significantly lower for both *K. pneumoniae* and *A. baumannii* compared to all
281 gonococcal datasets ($P < 0.0001$, **Fig 4d**).

282 While the SCM classifiers for CIP NS in *K. pneumoniae* performed significantly
283 better on the training sets than the testing sets (**Table S4**, $P < 0.0001$), indicating that
284 these models may be overfitted, there was no significant difference between RF-C model
285 performance on training and testing sets for either *K. pneumoniae* or *A. baumannii* ($P >$
286 0.194), suggesting that overfitting alone cannot explain the variable classifier
287 performance across different species. Down-sampling *K. pneumoniae* and *A. baumannii*
288 to match the CIP MIC distributions of the gonococcal datasets was infeasible due to the

289 narrow range of MICs tested for the former two species (**Table S7**). However, even after
290 down-sampling to equalize the number of S and NS strains within each dataset (**Table**
291 **S6, Fig. S3a-b**), performance of *K. pneumoniae* and *A. baumannii* CIP NS classifiers was
292 still significantly lower than that of gonococcal CIP NS classifiers, with the exception of
293 SCM classifiers based on the down-sampled *K. pneumoniae* dataset, which performed
294 roughly equivalently to SCM classifiers based on gonococcal datasets 2 and 6 ($P > 0.07$
295 for the SCM classifiers based on the down-sampled *K. pneumoniae* dataset compared to
296 SCM classifiers based on gonococcal datasets 2 and 6; $P < 0.0004$ for all other
297 comparisons, **Figure S3c**).

298 Direct association based on GyrA codon 83 mutations (equivalent to codon 91 in
299 gonococci) alone predicted CIP NS in *K. pneumoniae* with 86% sensitivity and 99%
300 specificity, and thus had a marginally higher bACC (92.5%) than for the SCM classifiers
301 and a substantially higher bACC than the RF classifiers. Similarly, for *A. baumannii*, GyrA
302 codon 81 mutations (equivalent to codon 91 in gonococci) alone predicted CIP NS in with
303 97% sensitivity and 98% specificity, and thus with a roughly equivalent bACC (97.5%) to
304 the SCM classifiers and a substantially higher bACC than the RF classifiers.

305

306

307 **Discussion**

308 ML offers an opportunity to leverage WGS data to aid in development of rapid molecular
309 diagnostics. While more comprehensive sampling of methods and parameters will be
310 necessary to optimize model performance, we demonstrate that multiple factors beyond
311 ML methods and parameters can affect model performance, reliability, and

312 interpretability. Our results affirmed that drugs associated with complex and/or diverse
313 resistance mechanisms present challenges to ML-based prediction of resistance
314 phenotypes and that sampling frame (*i.e.*, temporal range, geographic range, and/or
315 sampling approach) can substantially affect performance of such predictive models. We
316 demonstrated significant variability in performance and potential clinical utility of
317 predictive models based on different resistance metrics and further showed that the
318 capacity to model antibiotic resistance may be highly variable across different species.
319

320 **Variable performance of ML-based resistance prediction models by antibiotic**

321 Genotype-based resistance diagnostics have largely focused more on evaluating
322 the presence of resistance determinants and less on predicting the susceptibility profile
323 of a given isolate (8). However, in clinical settings where the empirical presumption is of
324 resistance, prediction that an isolate is susceptible to an antibiotic may be more important
325 in guiding treatment decisions. As such, the clinical utility of a genotype-based resistance
326 diagnostic may be determined by its capacity to accurately predict susceptibility
327 phenotype for multiple drugs.

328 While variable performance of ML-based predictive models has been observed
329 across different drugs (7, 8, 10, 11, 14, 15), it has often been attributed to dataset size
330 and/or imbalance (7, 14, 15). Further, while it is more difficult to predict resistance
331 phenotypes from genotypes for drugs that are associated with unknown, multifactorial,
332 and/or diverse resistance mechanisms than for drugs for which resistance can largely be
333 attributed to a single variant (14, 29), this caveat has been presented specifically as a
334 limitation of models based on known resistance loci in comparison to unbiased machine

335 learning-based MIC prediction using genome-wide feature sets (14). However, by
336 comparing performance of predictive models based on genome-wide feature sets
337 between CIP and AZM across multiple gonococcal datasets, we showed that even with
338 relatively large and phenotypically balanced datasets, ML algorithms cannot necessarily
339 be expected to successfully model complex and/or diverse resistance mechanisms,
340 particularly given that the representation of these resistance mechanisms in training
341 datasets is *a priori* unknown.

342 As a high proportion of reported AZM MICs in gonococci are within 1-2 doubling
343 dilutions of the NS breakpoints, it is possible that the inferior performance of AZM
344 classifiers is partly attributable to errors and/or variations in MIC testing. However, given
345 the noise of phenotypic MIC testing even with standardized protocols (32), this may be
346 an inherent limitation of NS classifiers when low-level resistance is common. Further,
347 while we show that removing strains with MICs ≤2 doubling dilutions from the breakpoints
348 improved AZM classifier performance compared to AZM models trained and tested on
349 the full dataset, performance of AZM classifiers trained and tested on this restricted
350 dataset was still significantly lower than that of CIP classifiers, suggesting that additional
351 drug-specific factors, such resistance mechanism diversity and/or complexity, can
352 constrain classifier performance.

353

354 **Impact of demographic, geographic, and timeframe sampling bias on ML model**
355 **predictions of antibiotic resistance**

356 Sampling bias presents a substantial challenge in any predictive modeling, and
357 sampling from limited patient demographics or during limited time periods may have

358 considerable effects on the distributions of resistance phenotypes and resistance
359 mechanisms (33, 34). For example, in TB, the RpoB I491F mutation that has been
360 associated with failure of commercial RIF resistance diagnostic assays, including the
361 GeneXpert MTB/RIF assay, reportedly accounted for <5% of TB RIF resistance in most
362 countries, but, in Swaziland was found to be present in up to 30% of MDR-TB (35).
363 Further, as the focus with statistical classifiers is building models from feature sets that
364 can accurately predict an outcome, rather than understanding the association between
365 each of the features and the outcome, potential confounding effects from factors such as
366 population structure (36-38) or correlations among resistance profiles of different drugs
367 (13) are rarely considered.

368 By comparing performance of AZM NS classifiers across multiple training and
369 testing sets, we showed significant variation in performance of classifiers trained on a
370 large and diverse global collection across testing sets from different sampling frames. In
371 some cases of imbalanced datasets, models trained on datasets with a more
372 comprehensive representation of resistance phenotypes improve prediction accuracy.
373 Our results further demonstrate that the direction of dataset imbalance (*i.e.*, the ratio of
374 NS to S strains) is significantly correlated with the direction of model performance (*i.e.*,
375 the ratio of sensitivity to specificity), suggesting that, for example, optimizing sensitivity of
376 predictive models for drugs with low prevalence of NS strains may require substantial
377 enrichment of NS strains and/or down-sampling of S strains. However, while differential
378 classifier performance among different datasets may be partially attributable to differential
379 MIC distributions, our results also show variable classifier performance between datasets
380 even in the case of identical MIC distributions (and sample size) and further suggest that

381 heavier sampling across more geographic regions cannot necessarily be expected to
382 significantly improve model performance, as models trained on the aggregate global
383 gonococcal dataset did not improve prediction accuracy for most datasets.

384 This, together with decreased performance when excluding isolates from the
385 dataset from which the isolates being tested were derived, suggests that factors such as
386 population-specific resistance mechanisms, genetic divergence at resistance loci, and/or
387 confounding effects may constrain model reliability across populations, particularly in the
388 case of drugs like AZM with complex and/or diverse resistance mechanisms, where a
389 substantial portion of the model may be overfit, or based on confounding factors or noise,
390 rather than biologically-meaningful resistance variants. Further, it should be noted that
391 MIC testing methods varied between some datasets (and between strains within dataset
392 5), and such variations may represent an additional confounding factor influencing
393 classifier performance. Thus, both incorporation of methods to correct for potentially
394 confounding factors, such as population structure, as have been introduced for genome-
395 wide associate studies [15-17], and increased availability of paired WGS and antibiotic
396 susceptibility data produced by consistent standardized protocols may improve reliability
397 of machine learning-based prediction of antibiotic resistance across different populations.
398

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

401 While measurement of MICs is vital for surveillance and investigation of resistance
402 mechanisms, resistance breakpoints that relate *in vitro* MIC measurements to expected
403 treatment outcomes inform clinical decision-making. However, standard breakpoints for

404 NS to a given drug in a given species are often informed less by treatment outcome data,
405 but rather factors such as pharmacokinetics and MIC distributions that can fail to account
406 for a variety of intra-host conditions that could influence drug efficacy (39-42). Recent
407 studies have shown that isolates that are classified as susceptible by standard
408 breakpoints but have higher MICs are associated with a greater risk of treatment failure
409 than isolates with lower MICs (43). Further, resistance breakpoints and testing protocols
410 can vary across different organizations, and thus incongruence across phenotypic
411 information included in the training data may introduce additional sources of error in
412 predictive modeling. By comparing performance of predictive models of CIP and AZM NS
413 based on EUCAST and CLSI breakpoints, we demonstrated breakpoint-specific
414 performance of models. For CIP, such breakpoint-specific performance is likely largely
415 attributable to variations in MIC testing protocols and thus ambiguous classification of
416 some strains by the EUCAST breakpoint. On the other hand, the substantially lower
417 performance of all AZM models based on the EUCAST breakpoint compared to those
418 based on the CLSI breakpoint suggests that many isolates with AZM MICs between the
419 two breakpoints lack genetic signatures that contribute to high model performance. While
420 the clinical relevance of AZM MICs between these two breakpoints in gonococci is
421 unclear, these isolates may be more likely to be associated with AZM treatment failure
422 than isolates with lower MICs, and thus evaluation of classifiers using only higher
423 breakpoints may misrepresent their diagnostic value, particularly in the absence of
424 sufficient treatment outcome data.

425 Models that predict MICs provide more refined output than a binary classifier but
426 generally achieve low rates of exact matches between phenotypic and predicted MICs

427 and even fairly variable 1-tier accuracies (14, 15, 29). Given the noise in phenotypic MIC
428 testing (32) and the potential lack of discriminating genetic features between isolates with
429 MICs separated by 1-2 doubling dilutions (14), MIC prediction models may be unlikely to
430 provide much better resolution than binary S/NS classifiers. Even if MIC predictions could
431 provide additional resolution, the most important criterion of such a diagnostic would likely
432 still be its ability to correctly predict resistance phenotypes relative to a clinically relevant
433 breakpoint. Thus, performance of MIC prediction models with respect to breakpoints may
434 be the biggest determinant of their diagnostic utility. By building MIC prediction models
435 for CIP and AZM in gonococci, we observed low rates of exact matches between
436 phenotypic and predicted MICs and variable 1-tier accuracies, with no relationship
437 between 1-tier accuracy and categorical agreement (*i.e.*, prediction accuracy relative to
438 NS breakpoints). Further, binary classifiers performed equivalently or better than MIC
439 prediction models.
440
441 **ML antibiotic resistance prediction model success varies across species**
442 Bacterial species with high genomic diversity (*e.g.*, open pangenomes) present
443 additional challenges to ML-based prediction of antibiotic resistance. Increased
444 resistance mechanism complexity and greater inter-isolate variation in resistance
445 mechanisms require more intensive sampling to capture a significant portion of the
446 resistome (47). On the technical side, even for heavily sampled species, when using
447 whole genome feature sets, the number of genetic features (*e.g.*, k-mers or SNPs) will
448 always be much larger than the number of observations (isolates), increasing the risk of
449 overfitting (a situation that arises with so-called 'fat data'; (12)). This raises concern in

450 species with open pangenomes, as the ratio of genetic features to the number of genomes
451 is larger and the number of unique genetic features per number of genomes does not
452 plateau. By comparing classifier performance in predicting CIP NS across gonococci, *K.*
453 *pneumoniae*, and *A. baumannii*, we show that classifiers generally did not perform as well
454 for species with open genomes (*K. pneumoniae* or *A. baumannii*) as for gonococci.
455 Further, while a single GyrA mutation could explain the majority of CIP NS across all
456 species evaluated here, unlike in gonococci and *A. baumannii* where this mutation
457 explained ≥97% of CIP NS, 14% of CIP NS in *K. pneumoniae* could not be explained by
458 this mutation, suggesting increased CIP resistance mechanism diversity and/or
459 complexity in this species. Increased sampling, different methods, and/or finer tuning of
460 hyperparameters may yield increased prediction accuracy for drug resistance in species
461 with open genomes. For example, Nguyen et al., 2018 reported a mean bACC of 98.5%
462 (average VME and ME rates of 0.5% and 2.5%, respectively) using a decision tree-based
463 extreme gradient boosting regression model to predict CIP MICs for the *K. pneumoniae*
464 strains assessed here (14), and adjusting for confounding factors such as population
465 structure or variation in MIC testing method may yield more consistent prediction
466 accuracies across species. However, our results demonstrate clear variation in potential
467 limitations of genotype-to-resistance-phenotype models across different species.
468
469 Given the biological and epidemiological disparities associated with resistance to
470 different drugs in different clinical populations and bacterial species, and their evident
471 impact on performance of predictive models, successful implementation of genotype-
472 based resistance diagnostics will likely require sustained comprehensive sampling to

473 ensure representation of complex, diverse, and/or novel resistance mechanisms,
474 customized modeling, and incorporation of feedback mechanisms based on treatment
475 outcome data. Further evaluation of additional ML methods and datasets may reveal
476 more quantitative requirements and limitations associated with the application of
477 genotype-to-resistance-phenotype predictive modeling in the clinical setting.
478

479 **Materials and Methods:**

480 **Isolate selection and dataset preparation**

481 See **Table 1** for details of the datasets assessed and **Table S7** for per-strain information.
482 All gonococcal datasets contained a minimum of 200 isolates with WGS (Illumina MiSeq,
483 HiSeq, or NextSeq) and MICs available for both CIP and AZM (by agar dilution and/or
484 Etest). Isolates lacking CIP and AZM MIC data were excluded. MIC testing methods
485 varied within datasets, as reported (10-13, 17, 18, 29).

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

490 MIC data were obtained from the associated publications, except in the cases of
491 dataset 1 (NCBI Bioproject PRJEB10016; see **Table S7**) and dataset 9, which were
492 obtained from the NCBI BioSample database (https://www.ncbi.nlm.nih.gov/biosample).
493 Raw sequence data were downloaded from the NCBI Sequence Read Archive
494 (https://www.ncbi.nlm.nih.gov/sra). Genomes were assembled using SPAdes (48) with
495 default parameters, and assembly quality was assessed using QUAST (49). Contigs <200

496 bp in length and/or with <10x coverage were removed. Isolates with assembly N50s below
497 two standard deviations of the dataset mean were removed.
498
499 **Evaluation of known resistance variants**
500 Previously identified genetic loci associated with reduced susceptibility to CIP or AZM in
501 gonococci are indicated in **Tables S1-S2**, respectively. The sequences of these loci were
502 extracted from the gonococcus genome assemblies using BLAST (50) followed by
503 MUSCLE alignment (51) to assess the presence or absence of known resistance variants.
504 The presence or absence of quinolone resistance determining mutations in *gyrA* was
505 similarly assessed in *K. pneumoniae* and *A. baumannii* assemblies. Presence or absence
506 of gonococcal AZM resistance mutations in the multi-copy 23S rRNA gene was assessed
507 using BWA-MEM(52) to map raw reads to a single 23S rRNA allele from the NCCP11945
508 reference isolate (NGK_rrna23s4), the Picard toolkit
509 (http://broadinstitute.github.io/picard) to identify duplicate reads, and Pilon (53) to
510 determine the mapping quality-weighted percentage of each nucleotide at the sites of
511 interest.
512
513 **ML-based prediction of resistance phenotypes**
514 Predictive modeling was carried out using SCM and RF algorithms, implemented in the
515 Kover (11, 12) and ranger (54) packages, respectively. K-mer profiles (abundance profiles
516 of all unique words of length k in each genome) were generated from the assembled
517 contigs using the DSK k-mer counting software (55) with k=31, a length commonly used
518 in bacterial genomic analysis (11, 12, 36, 56). For each dataset, 31-mer profiles for all
519 strains were combined using the combinekmers tool implemented in SEER (36),

520 removing 31-mers that were not present in more than one genome in the dataset. Final
521 matrices used for model training and prediction were generated by converting the
522 combined 31-mer counts for each dataset into presence/absence matrices. For each
523 SCM binary classification analysis (using S/NS phenotypes based on the two different
524 breakpoints for each drug), the best conjunctive and/or disjunctive model using a
525 maximum of five rules was selected using five-fold cross-validation, testing the suggested
526 broad range of values for the trade-off hyperparameter of 0.1, 0.178, 0.316, 0.562, 1.0,
527 1.778, 3.162, 5.623, 10.0, and 999999.0 to determine the optimal rule scoring function
528 (http://aldro61.github.io/kover/doc_learning.html). In order to assess binary classification
529 across multiple methods, RF was also used to build binary classifiers (RF-C) using S/NS
530 phenotypes. Further, to compare performance of binary classifiers to MIC prediction
531 models, RF was used to build multi-class classification (RF-mC) and regression (RF-R)
532 models based on $\log_2$(MIC) data. For all RF analyses, forests were grown to 1000 trees
533 using node impurity to assess variable importance and five-fold cross-validation to
534 determine the most appropriate hyperparameters (yielding the highest bACC or 1-tier
535 accuracy for NS- or MIC-based models, respectively), testing maximum tree depths of 5,
536 10, 100, and unlimited and mtry (number of features to split at each node) values of 1000,
537 10000, and either $\sqrt{p}$ or $p/3$, for classification and regression models, respectively, where
538 p is the total number of features (31-mers) in the dataset. While a grid search would
539 enable assessment of more combinations of different hyperparameter values and thus
540 finer tuning of hyperparameters, such an approach is computationally prohibitive on
541 datasets of this size. To standardize reported MIC ranges across datasets, CIP MICs
542 ≤0.008 μg/mL or ≥32 μg/mL were coded as 0.008 μg/mL or 32 μg/mL, respectively, and

543 AZM MICs ≤0.008 μg/mL or ≥32 μg/mL were coded as 0.03 μg/mL or 32 μg/mL,
544 respectively.

545 The set of SCM and RF analyses performed are indicated in **Tables S3** and **S6.**
546 For each of the seven individual gonococcal datasets, as well as the aggregate
547 gonococcal dataset (all gonococcal datasets combined, removing duplicate strains) and
548 the *K. pneumoniae* and *A. baumannii* datasets, training sets consisted of random sub-
549 samples of two-thirds of isolates from the dataset indicated (maintaining proportions of
550 each resistance phenotype from the original dataset), while the remaining isolates were
551 used to test performance of the model. Each set of analyses (for each combination of
552 dataset/drug/resistance metric/ML algorithm) was performed on 10 replicates, each with
553 a unique randomly partitioned training and testing set. For all gonococcal datasets,
554 separate models were trained and tested using the EUCAST (57) and CLSI (58)
555 breakpoints for NS to CIP. Four of the *N. gonorrhoeae* datasets had insufficient (<15) NS
556 isolates by the CLSI breakpoint for AZM non-susceptibility and thus were only assessed
557 at the EUCAST AZM breakpoint. CIP MICs for the *K. pneumoniae* isolates were not
558 available in the range of the EUCAST breakpoint (0.25 μg/mL), and thus only the CLSI
559 breakpoint for NS (>1 μg/mL) was assessed. For *A. baumannii*, the EUCAST and CLSI
560 breakpoints for ciprofloxacin NS are the same (>1 μg/mL). Due to the very limited range
561 of MICs within the BD-Phoenix testing thresholds and thus the CIP MICs available for *K.*
562 *pneumoniae* and *A. baumannii*, predictive models based on MICs were not generated for
563 these species. For analyses in **Table S6** where datasets were down-sampled to equalize
564 MIC distributions between datasets or the number of S and NS strains within datasets,

565 the required number of strains from the over-represented class(es) were selected at
566 random for removal.

567 Model performance was assessed by sensitivity (1 – VME rate), specificity (1 – ME
568 rate), and aggregate bACC (the average of the sensitivity and specificity (59)). bACC was
569 used as an aggregate measure of model performance as, unlike metrics such as raw
570 accuracy, error rate, and F1 score, it provides a balanced representation of false positive
571 and false negative rates, even in the case of dataset imbalance. For MIC prediction
572 models, the percentage of isolates with predicted MICs exactly matching the phenotypic
573 MICs (rounding to the nearest doubling dilution, in the case of regression models), as well
574 as the percentage of isolates with predicted MICs within one doubling dilution of
575 phenotypic MICs (1-tier accuracy), were also assessed. In order to account for variations
576 in MIC testing methods and thus in the dilutions assessed, criteria for exact match rates
577 and 1-tier accuracies were relaxed to include predictions within 0.5 doubling dilutions or
578 1.5 doubling dilutions, respectively, of the phenotypic MIC. Mean and 95% confidence
579 intervals for all metrics were calculated across the 10 replicates for each analysis.
580 Differential model performance between datasets or methods was evaluated by
581 comparing mean bACC between sets of replicates by two-tailed unpaired t-tests with
582 Welch's correction for unequal variance (α=0.05). Unless otherwise noted, all *P*-values
583 are derived from these unpaired t-tests. Relationships between MIC prediction accuracy
584 and bACC and between dataset imbalance and model performance were assessed by
585 Pearson correlation (α=0.05).

586

587

588 **Acknowledgements**

589 We thank Jung-Eun Shin, Mark Labrador, and members of the Grad Lab for helpful
590 discussion, and Julie Schillinger and Preeti Pathela for assistance identifying, selecting,
591 and characterizing the isolates from New York City. The authors declare no competing
592 interests.
593

594 **References:**

595 1. The Review on Antimicrobial Resistance. Tackling drug-resistant infections
596 globally: final report and recommendations. London, United Kingdom; 2016.
597 2. Zumla A, Al-Tawfiq JA, Enne VI, Kidd M, Drosten C, Breuer J, et al. Rapid point
598 of care diagnostic tests for viral and bacterial respiratory tract infections--needs,
599 advances, and future prospects. Lancet Infect Dis. 2014;14(11):1123-35.
600 3. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical
601 microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13(9):601-12.
602 4. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, et al.
603 Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug
604 susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis.
605 2015;15(10):1193-202.
606 5. Rigouts L, Gumusboga M, de Rijk WB, Nduwamahoro E, Uwizeye C, de Jong B,
607 et al. Rifampin resistance missed in automated liquid culture system for Mycobacterium
608 tuberculosis isolates with specific rpoB mutations. J Clin Microbiol. 2013;51(8):2641-5.
609 6. Mason A, Foster D, Bradley P, Golubchik T, Doumith M, Gordon NC, et al.
610 Accuracy of Different Bioinformatics Methods in Detecting Antibiotic Resistance and

611 Virulence Factors from Staphylococcus aureus Whole-Genome Sequences. J Clin
612 Microbiol. 2018;56(9).
613 7. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine
614 learning for classifying tuberculosis drug-resistance from DNA sequencing data.
615 Bioinformatics. 2018;34(10):1666-71.
616 8. Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham CD, et al.
617 Evaluation of Machine Learning and Rules-Based Approaches for Predicting
618 Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome
619 Sequence Data. Front Microbiol. 2016;7:1887.
620 9. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Jr., Walker H, et al. Validation of
621 beta-lactam minimum inhibitory concentration predictions for pneumococcal isolates
622 with newly encountered penicillin binding protein (PBP) sequences. BMC Genomics.
623 2017;18(1):621.
624 10. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid
625 antibiotic-resistance predictions from genome sequence data for Staphylococcus
626 aureus and Mycobacterium tuberculosis. Nat Commun. 2015;6:10063.
627 11. Drouin A, Giguere S, Deraspe M, Marchand M, Tyers M, Loo VG, et al.
628 Predictive computational phenotyping and biomarker discovery using reference-free
629 genome comparisons. BMC Genomics. 2016;17(1):754.
630 12. Drouin A, Letarte G, Raymond F, Marchand M, Corbeil J, Laviolette F.
631 Interpretable genotype-to-phenotype classifiers with performance guarantees. Sci Rep.
632 2019;9(1):4071.

633 13. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, et al. Antimicrobial
634 Resistance Prediction in PATRIC and RAST. Sci Rep. 2016;6:27930.
635 14. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing
636 an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci
637 Rep. 2018;8(1):421.
638 15. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al.
639 Using machine learning to predict antimicrobial minimum inhibitory concentrations and
640 associated genomic features for nontyphoidal Salmonella. J Clin Microbiol. 2018.
641 16. Santerre JW, Davis JJ, Xia F, Stevens R. Machine Learning for Antimicrobial
642 Resistance. arXiv e-prints. 2016.
643 17. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L.
644 Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data.
645 PLoS Comput Biol. 2018;14(12):e1006258.
646 18. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, et al. Prediction of
647 Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. J Clin
648 Microbiol. 2014;52(4):1182-91.
649 19. Marchland M, Shawe-Taylor J. The set covering machine. Journal of Machine
650 Learning Research. 2002;3:723-46.
651 20. Breiman L. Random forests. Machine Learning. 2001;45:5-32.
652 21. Hemarajata P, Yang S, Soge OO, Humphries RM, Klausner JD. Performance
653 and Verification of a Real-Time PCR Assay Targeting the gyrA Gene for Prediction of
654 Ciprofloxacin Resistance in Neisseria gonorrhoeae. J Clin Microbiol. 2016;54(3):805-8.

655 22. Siedner MJ, Pandori M, Castro L, Barry P, Whittington WL, Liska S, et al. Real-
656 time PCR assay for detection of quinolone-resistant Neisseria gonorrhoeae in urine
657 samples. J Clin Microbiol. 2007;45(4):1250-4.
658 23. Grad YH, Harris SR, Kirkcaldy RD, Green AG, Marks DS, Bentley SD, et al.
659 Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum
660 Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000-2013. J
661 Infect Dis. 2016;214(10):1579-87.
662 24. Wadsworth CB, Arnold BJ, Sater MRA, Grad YH. Azithromycin Resistance
663 through Interspecific Acquisition of an Epistasis-Dependent Efflux Pump Component
664 and Transcriptional Regulator in Neisseria gonorrhoeae. MBio. 2018;9(4).
665 25. Yakkala H, Samantarrai D, Gribskov M, Siddavattam D. Comparative genome
666 analysis reveals niche-specific genome expansion in Acinetobacter baumannii strains.
667 PLoS One. 2019;14(6):e0218204.
668 26. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al.
669 Genomic analysis of diversity, population structure, virulence, and antimicrobial
670 resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad
671 Sci U S A. 2015;112(27):E3574-81.
672 27. Harris SR, Cole MJ, Spiteri G, Sanchez-Buso L, Golparian D, Jacobsson S, et al.
673 Public health surveillance of multidrug-resistant clones of Neisseria gonorrhoeae in
674 Europe: a genomic survey. Lancet Infect Dis. 2018;18(7):758-68.
675 28. Yahara K, Nakayama SI, Shimuta K, Lee KI, Morita M, Kawahata T, et al.
676 Genomic surveillance of Neisseria gonorrhoeae to investigate the distribution and

677 evolution of antimicrobial-resistance determinants and lineages. Microb Genom.
678 2018;4(8).
679 29. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict
680 antibiotic MICs for Neisseria gonorrhoeae. J Antimicrob Chemother. 2017;72(7):1937-
681 47.
682 30. Demczuk W, Martin I, Peterson S, Bharat A, Van Domselaar G, Graham M, et al.
683 Genomic Epidemiology and Molecular Resistance Mechanisms of Azithromycin-
684 Resistant Neisseria gonorrhoeae in Canada from 1997 to 2014. J Clin Microbiol.
685 2016;54(5):1304-13.
686 31. Niehaus KE, Walker TM, Crook DW, Clifton TEAPA. Machine learning for the
687 prediction of antibacterial susceptibility in Mycobacterium tuberculosis. IEEE-EMBS
688 International Conference on Biomedical and Health Informatics (BHI)2014. p. 618-21.
689 32. Humphries RM, Ambler J, Mitchell SL, Castanheira M, Dingle T, Hindler JA, et al.
690 CLSI Methods Development and Standardization Working Group Best Practices for
691 Evaluation of Antimicrobial Susceptibility Tests. J Clin Microbiol. 2018;56(4).
692 33. Olesen SW, Torrone EA, Papp JR, Kirkcaldy RD, Lipsitch M, Grad YH.
693 Azithromycin susceptibility in Neisseria gonorrhoeae and seasonal macrolide use. J
694 Infect Dis. 2018;jiy551.
695 34. Unemo M, Shafer WM. Antibiotic resistance in Neisseria gonorrhoeae: origin,
696 evolution, and lessons learned for the future. Ann N Y Acad Sci. 2011;1230:E19-28.
697 35. Andre E, Goeminne L, Colmant A, Beckert P, Niemann S, Delmee M. Novel rapid
698 PCR for the detection of Ile491Phe rpoB mutation of Mycobacterium tuberculosis, a

699 rifampicin-resistance-conferring mutation undetected by commercial assays. Clin
700 Microbiol Infect. 2017;23(4):267 e5- e7.
701 36. Lees JA, Vehkala M, Valimaki N, Harris SR, Chewapreecha C, Croucher NJ, et
702 al. Sequence element enrichment analysis to determine the genetic basis of bacterial
703 phenotypes. Nat Commun. 2016;7:12797.
704 37. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for
705 association studies. Nat Genet. 2012;44(7):821-4.
706 38. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al.
707 Genomic analysis identifies targets of convergent positive selection in drug-resistant
708 Mycobacterium tuberculosis. Nat Genet. 2013;45(10):1183-9.
709 39. Prideaux B, Via LE, Zimmerman MD, Eum S, Sarathy J, O'Brien P, et al. The
710 association between sterilizing activity and drug distribution into tuberculosis lesions.
711 Nat Med. 2015;21(10):1223-7.
712 40. Tamma PD, Wu H, Gerber JS, Hsu AJ, Tekle T, Carroll KC, et al. Outcomes of
713 children with enterobacteriaceae bacteremia with reduced susceptibility to ceftriaxone:
714 do the revised breakpoints translate to improved patient outcomes? Pediatr Infect Dis J.
715 2013;32(9):965-9.
716 41. Bhat SV, Peleg AY, Lodise TP, Jr., Shutt KA, Capitano B, Potoski BA, et al.
717 Failure of current cefepime breakpoints to predict clinical outcomes of bacteremia
718 caused by gram-negative organisms. Antimicrob Agents Chemother. 2007;51(12):4390-
719 5.
720 42. Tam VH, Gamez EA, Weston JS, Gerard LN, Larocco MT, Caeiro JP, et al.
721 Outcomes of bacteremia due to Pseudomonas aeruginosa with reduced susceptibility to

722 piperacillin-tazobactam: implications on the appropriateness of the resistance
723 breakpoint. Clin Infect Dis. 2008;46(6):862-7.
724 43. Colangeli R, Jedrey H, Kim S, Connell R, Ma S, Chippada Venkata UD, et al.
725 Bacterial Factors That Predict Relapse after Tuberculosis Therapy. N Engl J Med.
726 2018;379(9):823-33.
727 44. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Deep
728 Learning Predicts Tuberculosis Drug Resistance Status from Whole-Genome
729 Sequencing Data. bioRxiv. 2018.
730 45. Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data--Recommendations
731 for the Use of Performance Metrics. 2013 Humaine Association Conference on
732 Affective Computing and Intelligent Interaction2013. p. 245-51.
733 46. US Food and Drug and Administration. Class II Special Controls Guidance
734 Document: Antimicrobial Susceptibility Test (AST) Systems. Rockville, MD; 2009.
735 47. Jeukens J, Freschi L, Kukavica-Ibrulj I, Emond-Rheault JG, Tucker NP,
736 Levesque RC. Genomics of antibiotic-resistance prediction in Pseudomonas
737 aeruginosa. Ann N Y Acad Sci. 2017.
738 48. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al.
739 SPAdes: a new genome assembly algorithm and its applications to single-cell
740 sequencing. J Comput Biol. 2012;19(5):455-77.
741 49. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
742 genome assemblies. Bioinformatics. 2013;29(8):1072-5.
743 50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment
744 search tool. J Mol Biol. 1990;215(3):403-10.

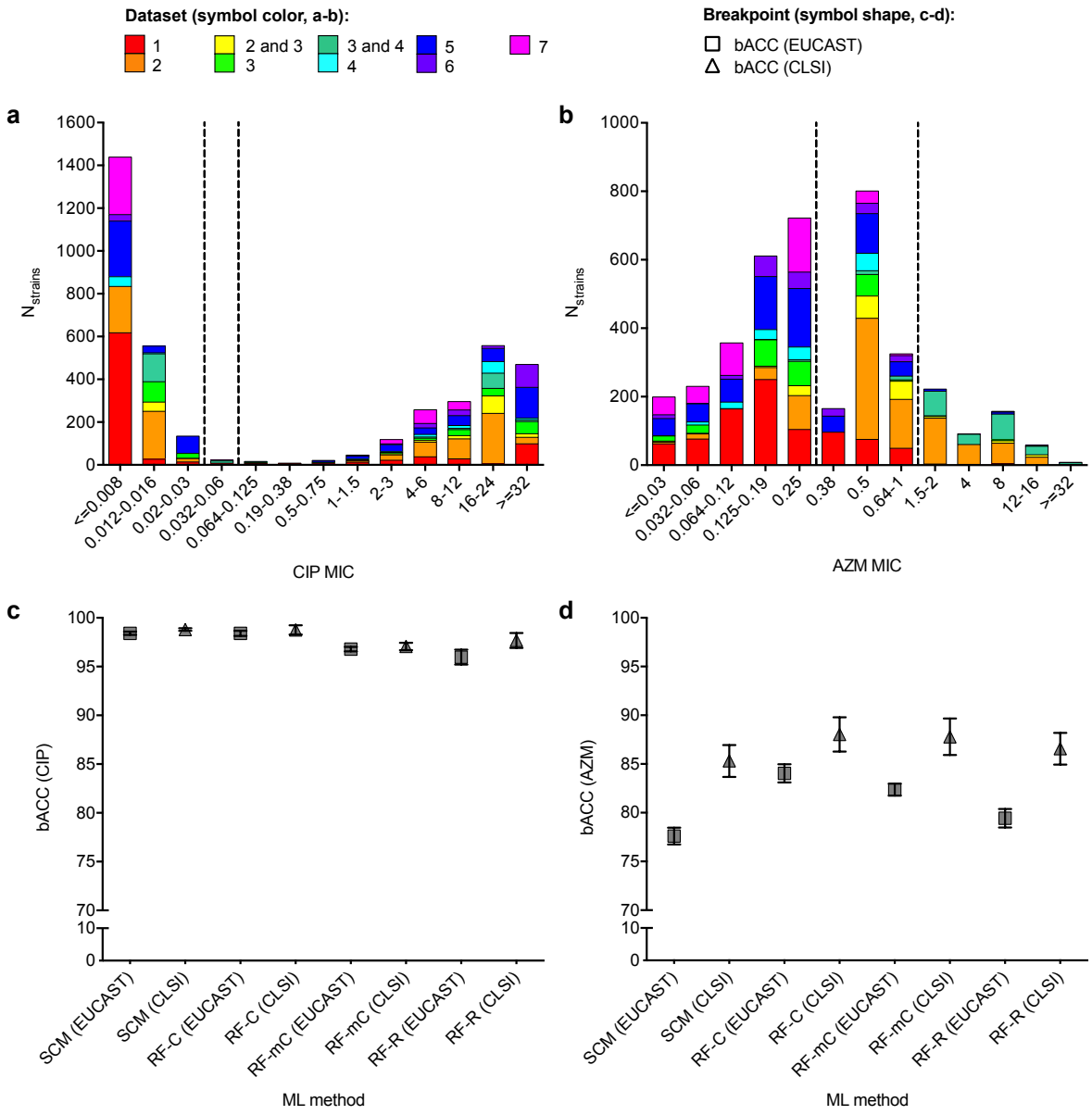
745 51. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
746 throughput. Nucleic Acids Res. 2004;32(5):1792-7.
747 52. Li H. Aligning sequence reads, clone sequences and assembly contigs with
748 BWA-MEM. arXiv e-prints. 2013.
749 53. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon:
750 an integrated tool for comprehensive microbial variant detection and genome assembly
751 improvement. PLoS One. 2014;9(11):e112963.
752 54. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High
753 Dimensional Data in C++ and R. Journal of Statistical Software. 2017;77:1-17.
754 55. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage.
755 Bioinformatics. 2013;29(5):652-3.
756 56. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al.
757 Identifying lineage effects when controlling for population structure improves power in
758 bacterial association studies. Nat Microbiol. 2016;1:16041.
759 57. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint
760 tables for interpretation of MICs and zone diameters. Version 8.1, 2018. [Available
761 from: http://www.eucast.org/.
762 58. Clinical and Laboratory Standards Institute. CLSI M100: Performance Standards
763 for Antimicrobial Susceptibility Testing, 29th Edition. 2019.
764 59. Bekkar M, Djemaa HK, Alitouche TA. Evaluation Measures for Models
765 Assessment over Imbalanced Data Sets. Journal of Information Engineering and
766 Applications. 2013;3(10):27-38.

767 60. De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, et al. Whole-
768 genome sequencing to determine transmission of Neisseria gonorrhoeae: an
769 observational study. Lancet Infect Dis. 2016;16(11):1295-303.
770 61. Demczuk W, Lynch T, Martin I, Van Domselaar G, Graham M, Bharat A, et al.
771 Whole-genome phylogenomic heterogeneity of Neisseria gonorrhoeae isolates with
772 decreased cephalosporin susceptibility collected in Canada between 1989 and 2013. J
773 Clin Microbiol. 2015;53(1):191-200.
774 62. Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, et al.
775 Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime
776 in the USA: a retrospective observational study. Lancet Infect Dis. 2014;14(3):220-6.
777 63. Lee RS, Seemann T, Heffernan H, Kwong JC, Goncalves da Silva A, Carter GP,
778 et al. Genomic epidemiology and antimicrobial resistance of Neisseria gonorrhoeae in
779 New Zealand. J Antimicrob Chemother. 2018;73(2):353-64.
780 64. Lesho EP, Waterman PE, Chukwuma U, McAuliffe K, Neumann C, Julius MD, et
781 al. The antimicrobial resistance monitoring and research (ARMoR) program: the US
782 Department of Defense response to escalating antimicrobial resistance. Clin Infect Dis.
783 2014;59(3):390-7.
784
785 **Table and figure legends:**
786
787 **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, ERP000144, ERP001405, ERP008891, SRP072971 (29)	671	2004-2014	International (UK, Canada, US)	Surveys from Brighton, UK (60) and nationwide sites in Canada (30, 61) and the US (23, 62); Canadian samples enriched for CRO and AZM resistance; US samples enriched for CFX resistance; US samples from male patients only
	4	SRP050190, SRP065041 (30, 61)	383	1989-2014	National (Canada)	Surveys from nationwide sites in Canada; enriched for CRO and AZM resistance
	5	ERP010312 (27)	714	2013	International (Europe)	Survey from clinics and hospitals across 21 European countries
	6	DRP004052 (28)	204	2015	National (Japan)	Survey from clinics in Kyoto and Osaka; male patients only
	7	SRP111927 (63)	398	2014-2015	National (New Zealand)	Survey from nationwide diagnostic labs
K. pneumoniae	8	SRP102664 (14)	1560	2011-2017	Houston, TX (US)	Survey from citywide hospital system; enriched for β-lactam resistance
A. baumannii	9	SRP065910 (64)	702	2000-2012	National (US)	Survey from clinics and hospitals within the US military healthcare system

788 CFX, cefixime; CRO, ceftriaxone; AZM, azithromycin
789



790

791 **Figure 1. Differential performance of machine learning-based prediction models for**
792 **ciprofloxacin and azithromycin resistance in gonococci**. Histograms showing the
793 distributions of **(a)** ciprofloxacin (CIP) and **(b)** azithromycin (AZM) minimum inhibitory
794 concentrations (MICs) in the gonococcal isolates assessed here. Bar color indicates the
795 study or studies associated with the isolates. Dashed lines indicate the **(a)** EUCAST and
796 CLSI breakpoints for non-susceptibility (NS, >0.03 μg/mL and >0.06 μg/mL, respectively)

797 for CIP and the **(b)** EUCAST and CLSI breakpoints for non-susceptibility (>0.25 μg/mL
798 and >1 μg/mL, respectively) for AZM. Note that there was some overlap in strains from
799 the US between datasets 2 and 3 and in strains from Canada between datasets 3 and 4;
800 such strains are indicated in **(a)** and **(b)** as belonging to datasets 2 and 3 and 3 and 4,
801 respectively. Mean balanced accuracy (bACC) with 95% confidence intervals of predictive
802 models for **(c)** CIP NS and **(d)** AZM NS trained and tested on the aggregate gonococcal
803 dataset. Symbol colors in (**a-b**) indicate the datasets from which the training and testing
804 sets were derived. Symbol shapes in (**c-d**) indicate the NS breakpoint. SCM, set covering
805 machine; RF-C, random forest classification; RF-mC, random forest multi-class
806 classification; RF-R, random forest regression.

807

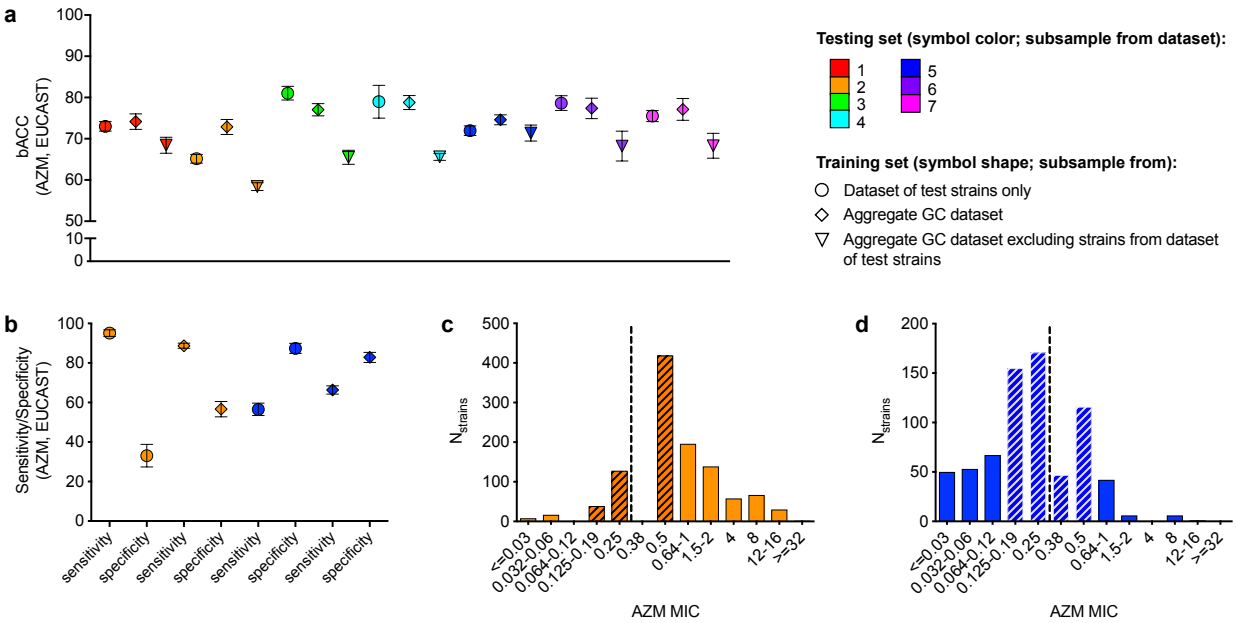


Figure 2. Differential performance of random forest classifiers across different datasets. (a) Mean balanced accuracy (bACC) with 95% confidence intervals of RF-C predictive models for gonococci (GC) azithromycin (AZM) non-susceptibility based on the EUCAST breakpoint. **(b)** Mean sensitivity and specificity with 95% confidence intervals of RF-C predictive models for GC AZM non-susceptibility in datasets 2 and 5. Histograms showing the distributions of AZM minimum inhibitory concentrations (MICs) in **(c)** dataset 2 and **(d)** dataset 5. Symbol colors in **(a)** and **(b)** indicate the dataset from which the testing set was derived, while symbol shape in **(a)** and **(b)** indicates the dataset from which the training set was derived. Hatching in **(c)** and **(d)** indicates MICs within one doubling dilution of the EUCAST breakpoint (designated by dashed lines).

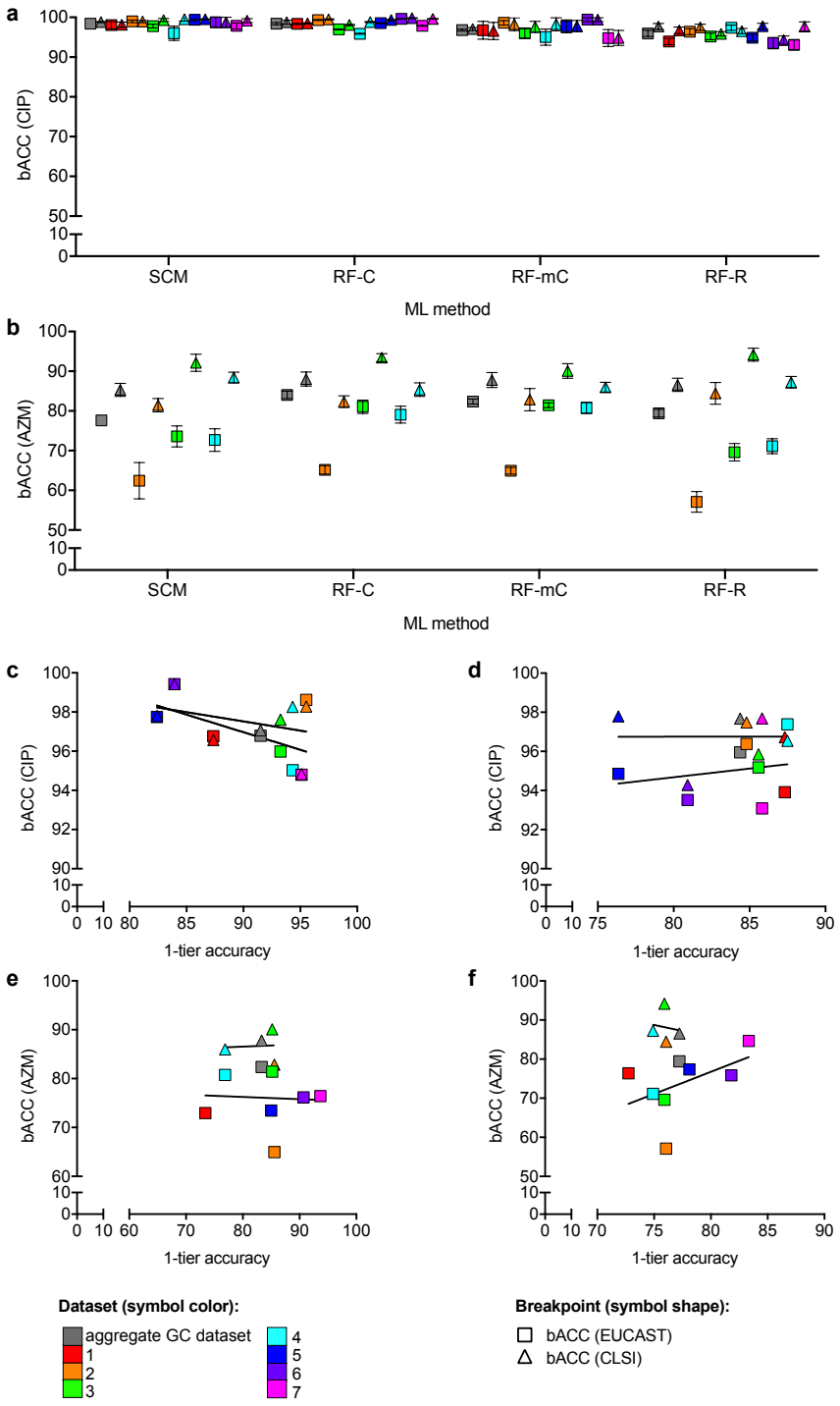
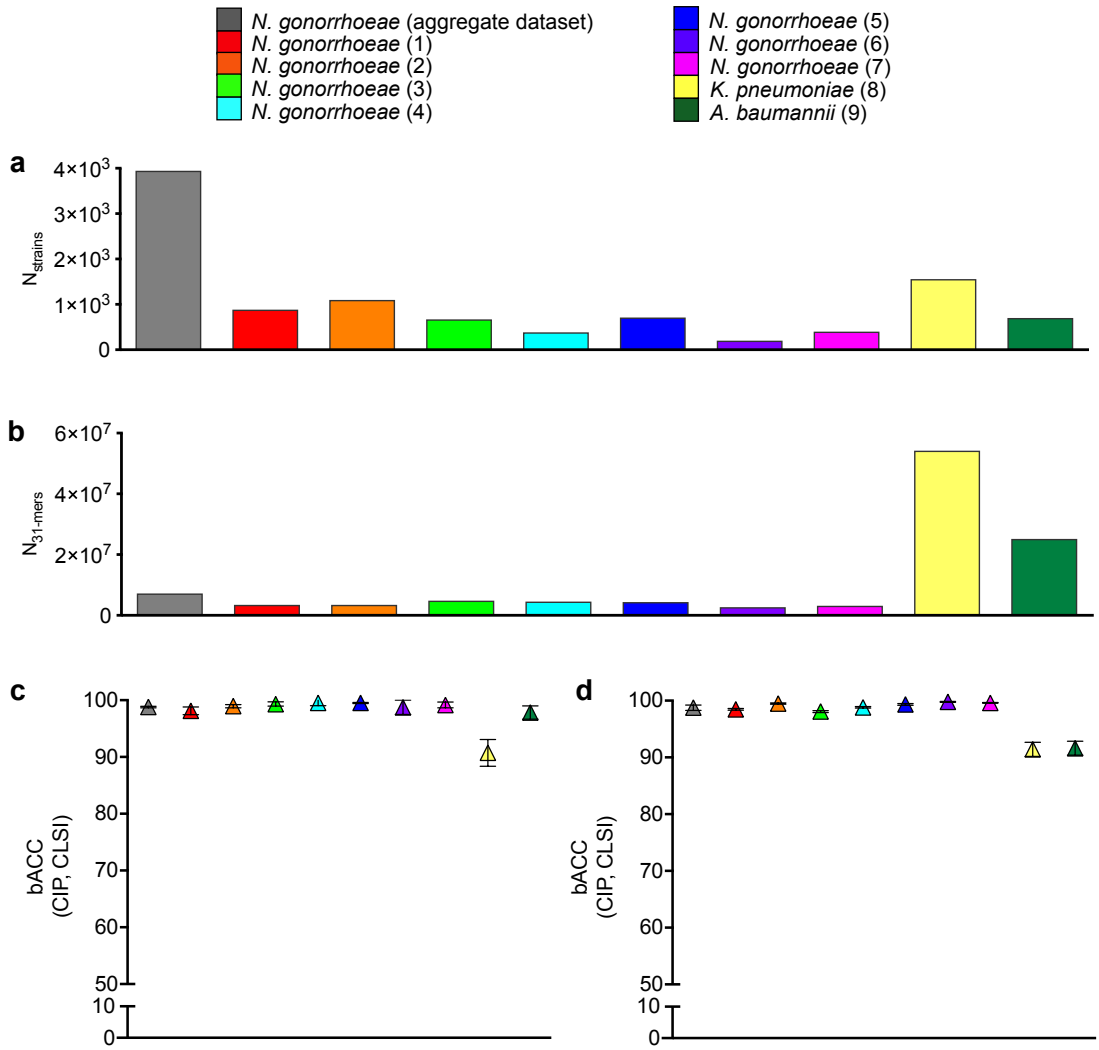


Figure 3. Differential performance of machine learning-based prediction models based on different resistance metrics in gonococci. Mean balanced accuracy (bACC) with 95% confidence intervals of predictive models for **(a)** ciprofloxacin non-susceptibility

824 (CIP NS) across all datasets and **(b)** azithromycin (AZM) NS for all datasets for which
825 both NS breakpoints were evaluated. Scatter plots comparing the mean 1-tier accuracy
826 to the mean bACC for each gonococcal dataset derived from **(c-d)** CIP and **(e-f)** AZM
827 minimum inhibitory concentration (MIC) prediction models by **(c,e)** random forest multi-
828 class classification and **d,f** random forest regression. Symbol colors in **(a-f)** indicate the
829 datasets from which the training and testing sets were derived. Symbol shapes in **(a-f)**
830 indicate the NS breakpoint. The line of best fit for each of the breakpoints is indicated in
831 **(c-f)**. SCM, set covering machine; RF-C, random forest binary classification; RF-mC,
832 random forest multi-class classification; RF-R, random forest regression.
833



834

835 **Figure 4. *K. pneumoniae* and *A. baumannii* datasets are associated with higher**
836 **genetic diversity and lower performance of resistance prediction models.** Number
837 of **(a)** strains and **(b)** unique 31-mers present in the genomes of at least two strains in
838 each dataset. Mean balanced accuracy (bACC) with 95% confidence intervals achieved
839 by **c)** set covering machine and **d)** random forest classification models for ciprofloxacin
840 (CIP) NS by the CLSI breakpoints across gonococcal, *K. pneumoniae,* and *A. baumannii*
841 datasets.

842

843 **Supplementary Tables and Figures**

844 **Table S1.** Genetic variants previously associated with ciprofloxacin resistance in *N.*
845 *gonorrhoeae.*

846 **Table S2.** Genetic variants previously associated with azithromycin resistance in *N.*
847 *gonorrhoeae*.

848 **Table S3.** Summary of approach in the primary set covering machine and random forest
849 analyses.

850 **Table S4.** Performance (mean with 95% confidence intervals) of predictive models for
851 ciprofloxacin resistance from the primary set covering machine and random forest
852 analyses.

853 **Table S5.** Performance (mean with 95% confidence intervals) of predictive models for
854 azithromycin resistance from the primary set covering machine and random forest
855 analyses.

856 **Table S6.** Summary of approach in the additional random forest analyses for assessment
857 of sampling bias.

858 **Table S7.** Study ID, machine learning dataset(s), antibiotic susceptibility testing (AST)
859 methods, azithromycin (AZM) and ciprofloxacin (CIP) minimum inhibitory concentrations
860 (MICs) for all strains assessed.

861 **Figure S1. MIC distribution influences classifier results but cannot explain all drug-**
862 **specific classifier performance.** Histograms showing azithromycin (AZM) minimum
863 inhibitory concentration (MIC) distributions for the aggregate gonococcal dataset after
864 down-sampling to remove all strains with MICs ≤2 doubling dilutions of the **(a)** EUCAST
865 or **(b)** CLSI breakpoint. **(c)** Mean balanced accuracy (bACC) with 95% confidence

866 intervals of SCM RF-C predictive models trained and tested on down-sampled aggregate
867 gonococcal datasets.

868 **Figure S2. Dataset imbalance influences classifier results but cannot explain all**
869 **dataset-specific classifier performance. (a)** Scatter plot showing the relationship
870 between the ratio of azithromycin (AZM) non-susceptible (NS) strains to susceptible (S)
871 strains (by the EUCAST breakpoint) in each dataset and the ratio of sensitivity to
872 specificity achieved by set covering machine (SCM) and random forest binary
873 classification (RF-C) methods. **(b)** Histogram showing the AZM minimum inhibitory
874 concentration (MIC) distribution for both datasets 2 and 4 after down-sampling to equalize
875 number of strains and MIC distributions between datasets. **(c)** Mean balanced accuracy
876 (bACC) with 95% confidence intervals of RF-C predictive AZM NS models trained and
877 tested on down-sampled datasets 2 and 4. Symbol colors in **(a)** indicated the machine
878 learning (ML) method. Symbol colors **(b)** indicate the down-sampled dataset from which
879 the training and testing sets were derived.

880 **Figure S3. Down-sampling to balance resistance phenotypes does ameliorate**
881 **cross-species variation in classifier performance.** Number of **(a)** strains and **(b)**
882 unique 31-mers present in the genomes of at least two strains in each dataset, after down-
883 sampling the *K. pneumoniae* and *A. baumannii* datasets to equalize the number of S and
884 NS strains within each dataset. Mean balanced accuracy (bACC) with 95% confidence
885 intervals achieved by **c)** set covering machine and **d)** random forest classification models
886 for ciprofloxacin (CIP) NS by the CLSI breakpoints across gonococcal, down-sampled *K.*
887 *pneumoniae,* and down-sampled *A. baumannii* datasets.