bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 **Metrics for Public Health Perspective Surveillance of Bacterial Antibiotic Resistance in Low-**
2 **and Middle-Income Countries**
3

4 Olga Tosas Auguet[1* ¥], Rene Niehus[2 ¥], Hyun Soon Gweon[3,4 ¥], James A. Berkley[1,5,6], Joseph
5 Waichungo[5], Tsi Njim[1], Jonathan D. Edgeworth[7], Rahul Batra[7], Kevin Chau[8], Jeremy Swann[8], Sarah
6 A. Walker[8,9], Tim E. A. Peto[8,9], Derrick W. Crook[8,9], Sarah Lamble[10], Paul Turner[1,11], Ben S.
7 Cooper[1,12], Nicole Stoesser[8,9]
8
9 [1] Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK
10 [2] Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA
11 [3] School of Biological Sciences, University of Reading, Reading, UK
12 [4] Centre for Ecology & Hydrology, Wallingford, UK
13 [5] KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya
14 [6] The Childhood Acute Illness and Nutrition (CHAIN) Network, Nairobi, Kenya
15 [7] Centre for Clinical Infection and Diagnostics Research (CIDR), Department of Infectious Diseases, King's
16 College London, London, UK
17 [8] Nuffield Department of Medicine, University of Oxford, Oxford, UK
18 [9] NIHR Health Protection Research Unit in Healthcare-associated Infections and Antimicrobial Resistance, UK
19 [10] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
20 [11] Cambodia-Oxford Medical Research Unit, Microbiology Department, Angkor Hospital for Children, Siem
21 Reap, Cambodia
22 [12] Mahidol–Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University,
23 Bangkok, Thailand
24
25
26 [*] Corresponding author
27 [¥] Joint first authors
28

29 **Abstract**

30 Antimicrobial resistance (AMR) is a global health threat, especially in low-/middle-income countries
31 (LMICs), where there is limited surveillance to inform empiric antibiotic treatment guidelines.
32 Enterobacterales are amongst the most important causes of drug-resistant bacterial infections. We
33 developed a novel AMR surveillance approach for Enterobacterales by profiling pooled human faecal
34 metagenomes from three sites (n=563 individuals; Cambodia, Kenya, UK) to derive a taxonomy-
35 adjusted AMR metric ("resistance potential") which could be used to predict the aggregate percentage
36 of resistant invasive Enterobacterales infections within each setting. Samples were sequenced
37 (Illumina); taxonomic and resistance gene profiling was performed using ResPipe. Data on organisms
38 causing bacteraemia and meningitis and antibiotic susceptibility test results from 2010-2017 were
39 collated for each site. Bayesian generalised linear models with a binomial likelihood were fitted to
40 determine the capacity of the resistance potential to predict AMR in Enterobacterales infections in

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

41 each setting. The most informative model accurately predicted the numbers of resistant infections in
42 the target populations for 14/14 of antibiotics in the UK, 12/12 in Kenya, and 9/12 in Cambodia.
43 Intermittent metagenomics of pooled human samples could represent a powerful pragmatic and
44 economical approach for determining and monitoring AMR in clinical infections, especially in
45 resource-limited settings.

46

47 Introduction

48 Antimicrobial resistance (AMR) is a global health emergency[1], and imposes a particularly large
49 socioeconomic burden in resource-limited settings, where bacterial infections and several other
50 drivers of AMR commonly co-occur and effective antibiotics may be unavailable or unaffordable[2]. A
51 key pillar in AMR mitigation is the development of effective and standardised AMR surveillance, to
52 monitor trends, inform empiric treatment guidelines, identify emerging AMR threats, and monitor the
53 impact of interventions. There has been significant investment in surveillance capacity, such as by the
54 UK's Fleming Fund, and an attempt to promote standardised collection, analysis and sharing of global
55 AMR data with an emphasis on capturing clinical and microbiological information, encapsulated in
56 the WHO Global Antimicrobial Resistance Surveillance System (GLASS)[3]. However, limitations in
57 implementing GLASS include the time taken to develop robust infrastructural capacity to support data
58 collection in regions where AMR is most relevant or prevalent, and the difficulty in obtaining
59 systematic datasets even from enrolled countries with adequate infrastructure, especially outside
60 tertiary or University centres. Surveillance strategies which could bridge or complement the
61 implementation of approaches such as GLASS would be helpful.

62

63 Colonisation with specific species and/or drug-resistant organisms, such as nasal colonisation with
64 *Staphylococcus aureus*[4], or rectal colonisation with carbapenemase-producing Enterobacterales[5], is
65 associated with risk of infection by these organisms. Metagenomic approaches are less biased than
66 targeted approaches which capture specific organism/resistance phenotypes of interest, and obviate
67 the need for culturing individual organisms. Resistance gene abundances and taxonomic distributions
68 in metagenomes are increasingly mined for a range of applications in the study of AMR, including as

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

69 correlates for national antibiotic exposures[6,7] in the case of human gut metagenomes, or as an
70 approach to monitoring global AMR in the case of sewage[8]. However, to our knowledge, no study to
71 date has used taxonomic and resistome profiles of pooled metagenomes to directly estimate the AMR
72 prevalence in clinical isolates within the same population, across a range of species and antimicrobial
73 classes. This approach would enable intermittent, strategic sampling of a subset of individuals in a
74 population to estimate the burden of AMR in clinical isolates, facilitating evidence-based
75 development of empiric treatment guidelines without the need for isolate-based microbiological
76 surveillance. Most samples taken to assay colonisation (e.g. faeces/rectal swabs, nasal/throat swabs)
77 are relatively non-invasive and acceptable for individuals, and tolerated by particularly vulnerable
78 groups, such as neonates.
79
80 The concept of a taxonomy-adjusted AMR metric or AMR resistance potential for a metagenome has
81 been described previously[6,9] as the average metagenome fraction encoding resistance genes for a
82 particular antibiotic or antibiotic group, across all bacteria in a sample that can potentially carry such
83 resistance genes, based on known taxonomic ranges for the resistance gene families. To model the
84 benefit of such a metric in predicting resistance in clinical isolates within a population, we took
85 pooled faecal samples from a sub-population of individuals (>100) in three disparate geographic
86 settings with varying AMR prevalence, namely Cambodia, Kenya and the United Kingdom (UK), and
87 validated the model predictions using microbiological data from clinical isolates processed by
88 laboratories in these locations over a seven-year period (2010-2017).
89

90 Materials and Methods

91

92 *Samples and Settings*

93 Faecal material stored in three existing biobanks was chosen for study; ethical approval for the
94 broader use of these samples was in place. Samples comprised: (i) rectal swabs from children aged 1-
95 59 months with and without malnutrition, taken on admission to Kilifi County Hospital in Kilifi,
96 Kenya, from 1st April to 30th September 2016, and stored in Amies transport media + 1ml phosphate

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

97 buffered saline at -80°C ("Pharmacokinetics of Antimicrobials and Carriage of Antimicrobial
98 Resistance amongst Hospitalised Children with Severe Acute Malnutrition (FLACSAM)' study[10]
99 [KEMRI/SERU/CGMR- C/023/3161; OXTREC 47-15]); (ii) faecal samples taken from newborns on
100 admission to Angkor Hospital for Children in Siem Reap, Cambodia, from 11th September 2013 to
101 10th September 2014, and stored in tryptone soya broth + 10% glycerol at -80°C[11](OxTREC ref 1047-
102 13; this collection also included longitudinal samples taken from a subset of newborns during their
103 inpatient stay for another study); and (iii), rectal swabs (Eswab, Copan diagnostics, Murrieta, CA,
104 USA); 1ml Amies transport media) from individuals aged ≥18 years attending pre-admission clinics
105 or on admission to Guy's and St Thomas' NHS Foundation Trust, London, UK, between February
106 and May 2015, and stored at -80°C[12–14] ([REC: 14/LO/2085]). Rectal swabs and faecal samples have
107 both been used as approaches for surveying intestinal microbiota[15,16], and are thought to give similar
108 results[17].

109

110 For each study site, metadata associated with microbiology tests performed on blood and
111 cerebrospinal fluid samples (as most robustly representative of true causative pathogens) collected
112 within 0-72 hours of admission from 01/Jan/2010-31/May/2017 were collated. Each site has a
113 microbiology laboratory participating in external quality assurance schemes (e.g. UK National
114 External Quality Assessment Service, NEQAS) and is additionally accredited to UK ISO15189
115 (London laboratory) or WHO Good Clinical Laboratory Practice standards (Kilifi laboratory).
116 Catchment areas served by each laboratory vary: For Cambodia about two-thirds of the patients come
117 from within Siem Reap province[18,19]; in Kenya the population served is mostly rural, within the
118 coastal Kilifi District[20]; and in London the laboratory largely serves a South London community of
119 approximately 0.5 million people and also regularly provides services to international patients and
120 patients from other sites in the UK[21]. Collated metadata included bacterial species identification
121 results, available antibiotic susceptibility testing (AST) results, specimen type and basic patient details
122 to validate aggregate-level stratification by age. Samples were processed using standard operating
123 procedures in accordance with accredited guidelines. In the UK, the VITEK system (bioMérieux,
124 Marcy-l'Etoile, France) was used for AST and performed according to the British Society for

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

125 Antimicrobial Chemotherapy standards[22] (BSAC). In Cambodia and Kenya, AST was performed
126 using a standardised disk diffusion method following the Clinical and Laboratory Standards Institute
127 (CLSI) guidelines[23]. Where accurate AST results could not be achieved by simple disk diffusion,
128 minimum inhibitory concentrations (MICs) were determined by Etest in both settings. The infection
129 metadata was collated for infants < 90 days of age in Cambodia, ≤ 60 months of age in Kenya and ≥
130 18 years of age in the UK.
131

132 ***DNA Extraction***

133 Samples from Cambodia and Kenya were shipped to the Nuffield Department of Medicine
134 (University of Oxford, UK) for extraction; extractions for London samples took place at the Centre
135 for Clinical Infection and Diagnostics Research (CIDR-King's College London). DNA was extracted
136 from each sample using the MoBio PowerSoil® DNA isolation kit (Qiagen, Hilden, Germany), as per
137 the manufacturer's instructions with optimisation steps to achieve sufficient DNA yields for
138 sequencing (ideally ≥300ng DNA/34ul, with a view to obtaining ≥20Gbp (Giga base pairs) of data per
139 sample). See Supplementary Methods 1 & 2. Known copy numbers of internal standards consisting of
140 *Thermus thermophilus* HB8 genomic DNA[24] (not normally present in faecal samples) were added to
141 each sample prior to the addition of Solution C1 (i.e. 8.75 ul per sample [1ng/ul of Thermus DNA]).
142 The presence of *T. thermophilus* was ascertained following sequencing by mapping reads to the
143 Thermus reference genome.
144

145 ***Sample Pooling***

146 DNA extracts were stored at -20°C and then pooled and sequenced at the Wellcome Trust Centre for
147 Human Genetics, Oxford, UK. For each study site, we created a "population pool", which consisted of
148 the pooling of equimolar concentrations of all extracts from that setting with ≥1ng DNA/µl. To
149 validate our pooling approach, we also created one smaller pool in each setting, a so-called "30-
150 sample pool", which consisted of equimolar concentrations of 30 randomly selected extracts with
151 ≥300ng DNA/34µl. An aliquot from each extract included in 30-sample pools was in turn sequenced
152 individually for the validation study (i.e. sequenced extracts from 90 individuals in total). An aliquot

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

153 from all extracts sequenced individually and included in the 30-sample pools was also included in
154 population pools.
155
156 ***Metagenomic Sequencing***
157 Sequencing of all samples (pools and individual extracts) was performed using the HiSeq 4000
158 Illumina platform, generating 150bp paired-end reads (i.e. 96 metagenomes [n=90 individual
159 metagenomes, n=3 30-sample pools, n=3 population pools]). 500ng of DNA from each sample was
160 used for library preparation. Libraries were constructed using the NEBNext Ultra DNA Sample Prep
161 Master Mix Kit (NEB) with minor modifications and a custom automated protocol on a Biomek FX
162 (Beckman)[25]. At the time of sequencing, the HiSeq 4000 produced on average 72-90 Gbp of data per
163 lane. We sequenced four individual extracts per lane to obtain on average 18-22.5 Gbp of data per
164 sample. For the pooled samples, we sequenced one 30-sample-pool plus one population-pool per lane
165 to obtain on average 36-45 Gbp of data per pool. Metagenomic data was obtained once for each
166 distinct sample or pool; there were no technical replicates due to the expense of high-throughput
167 sequencing.
168
169 ***Sequence Data Processing***
170 We determined the taxonomic abundance of bacterial species and resistance genes at individual and
171 pooled sample levels using a recently developed bioinformatics pipeline[26]. This pipeline incorporated
172 established approaches to taxonomic profiling, and an adapted approach to quantify resistance gene
173 markers present in a metagenome (for details of the method, see[26]). Briefly, the sequenced paired-end
174 reads were quality-filtered based on PHRED scores (≥ Q25 and ≥ 50 bp), and adapters removed using
175 TrimGalore[27]. For profiling the abundance of bacterial species, the quality-filtered sequences were
176 classified with Kraken2[28] (v.2.0.8-beta) against bacteria, plasmid, viral and human genome sequences
177 recovered (12 July 2019) from NCBI. With the taxonomic classification from Kraken2 and
178 information about species specific versus non-specific genetic regions we estimated true abundance at
179 the species level using Bracken[29] (v.2.5.0), which was subsequently used for deriving total aggregate
180 counts of bacterial taxa. For profiling resistance genes, the quality-filtered sequences were mapped

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

181 against the Comprehensive Antibiotic Resistance Database[30,31] (CARD, v.3.0.3) using BBMAP[32]
182 (v.37.72) at 100% sequence identity. The number of sequences that mapped to each resistance gene
183 were subsequently corrected to remove resistance gene length bias. This was done using four metrics,
184 namely (1) specific read count (number of sequences that map exclusively to the resistance gene); (2)
185 specific lateral coverage (proportion of the resistance gene covered by sequences mapping exclusively
186 to the gene); (3) resistance gene length; and (4) and average read length (average length of reads that
187 mapped to the resistance gene), and by the following formula: corrected gene count (CGC) = (specific
188 read count x average read length) / (resistance gene length x specific lateral coverage).
189
190 The CARD database attempts to classify each resistance gene variant by its association with AMR. To
191 be included in CARD, an AMR determinant must be described in a peer-reviewed scientific
192 publication, have its DNA sequence available in GenBank, and include clear experimental evidence of
193 elevated MIC over controls[31]. We used these data to map and aggregate counts of resistance
194 genes/variants associated with resistance to a specific antibiotic. In the process, we ranked the
195 resistance genes/variants into two categories, reflecting to some extent the public health risks posed[33],
196 and thereby creating two sets of antibiotic resistance gene metrics. The first (AMR_{DEF}; Supplementary
197 Data 1), included only AMR determinants with the "*Confers_Resistance_to_Antibiotic*" relationship
198 ontology term, whereby the gene associated with demonstrably elevated MIC is known to confer or
199 contribute to clinically relevant resistance to a specific antibiotic drug[31]. The second (AMR_{ALL};
200 Supplementary Data 2), contained corrected counts of all resistance genes with clear experimental
201 evidence of increasing the MIC, including those associated with clinically relevant resistance (as for
202 AMR_{DEF}), plus those without the "*Confers_Resistance_to_Antibiotic*" relationship ontology term. For
203 the purposes of this study we have used the term "resistance gene" to define any relevant genetic
204 marker of resistance, including genes that confer resistance by mutation (but can have a susceptible
205 wild type), and genes that confer resistance through presence/absence.
206
207 ***Validation of Pooling***

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

208 We evaluated to what extent pooled resistome data was a non-biased representation of the individual
209 resistomes making up the pool. Resistance gene abundances of the 30-sample pools and individually
210 sequenced samples were converted to relative abundances, such that gene abundances in each sample
211 summed to one. Then, for each of the three different settings, individual samples were used to
212 compute the empirical distribution of each gene by repeated random sampling of its relative gene
213 abundance out of the individual samples (bootstrapping with n=100,000 repeats). We were then able
214 to compare the pool abundance of each gene with its empirical distribution in the same setting
215 (within-setting comparison) and in the other two settings (across-setting comparison). We computed
216 the fraction of resistance genes for which the pool estimate was within 90% central quantile of the
217 empirical distribution. The resulting metric was restricted between 0 (i.e. 0% of resistance genes in
218 the pool were as expected given the individual resistomes) and 1 (i.e. 100% of resistance genes in the
219 pool were as expected). Because bootstrapping of gene abundances relies on having a sufficient
220 number of samples with non-zero abundance, we limited our analysis to genes present in ≥50% of all
221 individual samples (n=121 genes). Given the central quantile choice above (i.e. 90%), a value of
222 ~0.90 would imply a non-biased representation of individual resistomes by the pooled resistome. For
223 visualization, non-metric multidimensional scaling, an ordination-based method, was used to show
224 pair-wise dissimilarities between resistomes from population pools, 30-sample-pools and individual
225 sample means within and across settings. Individual sample means for each setting were, for each
226 AMR gene, the sum of CGCs across all individually sequenced samples.
227
228 ***Taxonomy Adjusted Resistance Potential Metrics***
229 We developed several candidate metrics of resistant infection risk, based on pool metagenomic data
230 on resistance gene abundance and bacterial species composition, and evaluated their potential to
231 accurately predict the likelihood of antibiotic resistant invasive infections in a population. We refer to
232 these as 'taxonomy-adjusted resistance potential (RP)' metrics, which consisted of two parameters.
233 The first parameter, R_{CGC}, was given through the sum of corrected gene counts (CGC) of variants
234 associated with resistance to a given antibiotic, j (R_{CGCj}) divided by the total CGC of all resistance
235 genes in the pool. $R_{CGC,j}$ was calculated based on either variants with experimental evidence of

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

236 increasing the MIC (AMR_{ALL}) or only variants known to confer clinically relevant resistance
237 (AMR_{DEF}). The second parameter, R_{Tax}, was given through the estimated abundance of a clinically
238 relevant bacterial grouping (derived from Bracken estimates) divided by the total estimated
239 abundance of bacterial taxa in the pool. The bacterial groupings tested were the Enterobacterales
240 order, Enterobacteriaceae family, and the grouping of the four most common and clinically relevant
241 bacterial genera/species within the Enterobacteriaceae family across sites (namely *Escherichia coli*,
242 *Klebsiella pneumoniae*, *Salmonella* spp, *Enterobacter* spp).
243
244 ***Bayesian Modelling***
245 With each taxonomy-adjusted RP, we fitted a Bayesian generalized linear model to the data and
246 applied model comparison. This allowed us to assess the potential of the different metrics to predict
247 observed antibiotic resistance amongst clinical invasive Enterobacterales isolates. We used de-
248 duplicated counts of isolates (unique bacterial species per antibiogram and patient-ID) for the
249 analyses. We let i denote the setting (Cambodia, Kenya or UK), and j the antibiotic (see below for a
250 list). We assumed that counts of resistant samples follow a binomial distribution. Our model then
251 predicts the count of resistance ($r_{i,j}$) among tested Enterobacterales isolates ($n_{i,j}$) using a probability of
252 resistance ($p_{i,j}$), which is modelled as
253
254 $$r_{i,j} \sim Binomial(p_{i,j}, n_{i,j})$$
255 $$logit(p_{i,j}) = \alpha_j + \beta_{1,j} R_{CGC,i,j} + \beta_{2,j} R_{Tax,i} \quad \text{(Equation 1)}$$
256
257 The model intercept (α) is specific for each antibiotic (j) but not setting (i), representing a baseline
258 propensity of resistance for any given antibiotic. Because resistance propensities can vary widely
259 between different antibiotics. We assume independent baselines (fixed effects). The setting-specific
260 information is $R_{Tax,i}$, which gives information about pathogen levels in setting i, as well as $R_{CGC,i,j}$,
261 which carries information about resistance toward antibiotic j in setting i. For $\beta_{1,j}$ and $\beta_{2,j}$, the
262 predictive effects of R_{CGC} and R_{Tax}, we assumed these to represent the clinical ecology of resistance

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

263 genes so that they are specific to each antibiotic, j, but not to each setting, i. We further assumed that
264 different antibiotics have different but related β-values (variable effects, specified below). We
265 included only those antibiotics that had existing antibiotic susceptibility test (AST) data in at least two
266 out of three settings (trimethoprim-sulfamethoxazole, nitrofurantoin, nalidixic acid, meropenem,
267 imipenem, gentamicin, ciprofloxacin, chloramphenicol, cefuroxime, ceftriaxone, ceftazidime,
268 cefpodoxime, cefoxitin, cefotaxime, ampicillin, amikacin); missing observations were excluded from
269 the likelihood evaluation. We fitted the above model with R_{CGC} being either AMR_{DEF} or AMR_{ALL} and
270 with R_{Tax} being either of the three bacterial groupings discussed earlier, yielding a total of six separate
271 model fits. Due to the limited number of infection isolates with AST results (especially in Cambodia),
272 we chose standard weakly informative priors for the intercept and the effect parameters. In addition,
273 we restricted the effect of gene abundance to be positive, reflecting our view that only a positive
274 association of resistance genes and clinical resistance is biologically reasonable. We therefore chose

275

276 $$\beta_{1,j} \sim N^{+}(\mu_{\beta,1}, \sigma_{\beta,1})$$

277 $$\beta_{2,j} \sim N(\mu_{\beta,2}, \sigma_{\beta,2})$$

278 $$\alpha_j \sim N(0,1)$$

279 $$\mu_{\beta,1} \sim N^{+}(0,1)$$

280 $$\mu_{\beta,2} \sim N(0,1)$$

281 $$\sigma_{\beta,1} \sim N^{+}(0,1)$$

282 $$\sigma_{\beta,2} \sim N^{+}(0,1)$$

283

284 where N denotes a normal distribution and N^{+} denotes a half-normal distribution covering only
285 positive values. Each model was fit using Stan software[34] (v2.19.1), with which we sampled 50,000
286 samples after a burn-in period of 5,000 samples using four independent chains.

287

288 The best taxonomy-adjusted RP metric was selected using Bayesian leave-one-out cross validation[35]
289 which estimates a model's pointwise out of sample prediction accuracy. The prediction accuracies are

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

290 then used to directly compare all models using stacking weights[36]. In brief, models with smaller cross-
291 validation errors (e.g. smaller prediction errors), get more weight relative to other models in the model
292 comparison. We also included in the comparison two models with R_{CGC} (either AMR_{DEF} or AMR_{ALL}),
293 following Equation (1), but without R_{Tax}. Finally, the overall value of using any taxonomy-adjusted
294 RP metric for predicting clinical resistance was assessed by including in the model comparison a
295 baseline model without predictors. The prediction accuracy of taxonomy-adjusted RP was also
296 assessed visually by comparing the best model's predictions of sample counts of resistance (and their
297 95% credible intervals [CI]) against the observed counts (Figure 5). For settings and antibiotics where
298 zero samples were tested, we imputed the sample size by computing the rounded mean of the sample
299 sizes of the other two settings. Model comparisons and all further data analyses were performed in R-
300 3.6.1 statistical software[37]. The dataset used for the Bayesian modelling is given in Supplementary
301 Data 3.
302
303 **Results**
304 The study included 210 admission samples from Kenya, 200 from the UK and 153 from Cambodia
305 (n=154 – 1 rejected sample), totalling 563 samples for metagenomic analysis (Fig 1). In addition, 76
306 follow-up samples were taken from 37/154 newborns in Cambodia during their inpatient stay or upon
307 hospital discharge for a separate project; these were processed alongside the study samples (Fig 1).
308 We only considered DNA extracts with yields ≥1ng/ul (79-89% of samples; Fig 1), and 19 DNA
309 extracts from the separate longitudinal study were included in the Cambodia population pool due to
310 processing error. In total, population pools in Kenya, the UK and Cambodia, comprised 177, 157, and
311 156 pooled sample extracts. Thirty high DNA-yield samples (≥9ng/ul) from each setting were used
312 for the validation study, as well as being included in the population pools. To prevent bias, potential
313 associations between high-yield samples and population traits were ruled out in advance. The total
314 Gbp of data obtained per population pool were 51.6 (Kenya), 55.1 (UK) and 52.6 (Cambodia). The
315 median Gbp obtained for individually sequenced samples were 24.2 (Kenya), 22.1 (Cambodia), 22.4
316 (UK).
317

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

318 **Fig 1. Sample Processing Workflow**

319

320

321 We identified 863 different antimicrobial resistance genes across any sample or pool (Cambodia: 684;
322 Kenya = 527; UK = 520), which were proven to increase the MIC for 163 antimicrobials (AMR_{ALL})
323 and known to confer clinically relevant resistance for 113 antimicrobials (AMR_{DEF}). The number of
324 resistance genes identified in population pools was largest in Cambodia (n=490), followed by the UK
325 (n=389) and Kenya (n=386). The median number of resistance gene types identified per individual
326 sample was also higher in Cambodia (median=162; IQR= 126-187 [Min-Max =33-231]), followed by
327 in Kenya (median=143; IQR= 127-205 [Min-Max =97-256]) and UK (median=134; IQR= 126-148;
328 [Min-Max =61-217]).

329

330 A summary of the Enterobacterales taxa identified from population pools and invasive infections in
331 each setting is given in Fig 2. Enterobacterales were the main bacterial taxa identified from population
332 pools in the UK (75.7%) and Cambodia (69.7%) but not in Kenya (32.4%) (Fig 2A). Within the
333 Enterobacterales, >95% of the bacteria were from the Enterobacteriaceae family in all settings (UK:
334 96.3%; Cambodia: 99.4%; Kenya: 99.1%). The predominant species within the Enterobacterales order
335 in population pools were *E. coli* and *K. pneumoniae*, followed by *Enterobacter* spp. (Fig 2E). These
336 species and genera combined accounted for 92.4% of all Enterobacterales taxa in Kenya, 88.5% in the
337 UK and 88.1% in Cambodia. The abundance of *E. coli*, was >20-fold higher than that of *K.*
338 *pneumoniae* in population pools from the UK (*E. coli*: 63.2%; *K. pneumoniae*: 2.2%) and Kenya (*E.*
339 *coli*: 28.4%; *K. pneumoniae*: 1.3%). In contrast both species had similar abundance in the Cambodia
340 population pool (*E. coli*: 30%; *K. pneumoniae*: 26.9%). *Enterobacter* spp. abundance was also higher
341 in Cambodia (4.5%) compared to the UK (1.6%) or Kenya (0.2%). The remaining Enterobacterales
342 comprised other genera, each being <2% of the total bacterial taxa in the three settings (Fig 2G).
343 Infections by Enterobacterales accounted for approximately a third of all blood and cerebrospinal
344 fluid infections in the three settings (Kenya: 36.8%; Cambodia: 33.0%; UK: 28.2%) (Fig 2B). Similar
345 to the findings from population pools, most of these Enterobacterales infections involved the

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

346 Enterobacteriaceae family (UK: 91.2%, Cambodia: 89.2%; Kenya: 91.8%; Fig 2B). Likewise, the
347 predominant Enterobacterales species in all settings were *E. coli* and *K. pneumoniae*, with the
348 proportion of *E. coli* infections being at least double that of *K. pneumoniae* in the UK (*E. coli*: 16.1%;
349 *K. pneumoniae*: 4.8%) and Kenya (*E. coli*: 13.8%; *K. pneumoniae*: 5.9%), but not in Cambodia (*E.*
350 *coli*: 13.7%; *K. pneumoniae*: 11.7%) (Fig 2F). *Enterobacter* spp. was the next most common
351 Enterobacterales genus in all settings (Cambodia: 3.1%; Kenya: 2.7%; UK: 2.2%), but the remaining
352 Enterobacterales species and genera accounted for <2% of the total invasive infections by any
353 bacterial order each in all three settings (Fig 2H). A notable exception was *Salmonella* spp., which
354 accounted for 9.9% of the total infections in Kenya (therefore also included in Fig 2F and the
355 equivalent plot for population pools [Fig 2E]). Details of all invasive infections by bacteria other than
356 the Enterobacterales are given in Supplementary Fig 1.
357

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

358 **Fig 2. Bacterial (Enterobacterales) taxa in population pools and in blood and cerebrospinal fluid**
359 **infections from Cambodia, Kenya and UK.**

360 Panels for population pools (A, C, E, G) show, for each setting, the abundances of Enterobacterales taxa divided
361 by the total abundance of bacterial taxa in a pool. Abundances are derived from Bracken estimates. Panels for
362 invasive infection data (B, D, F, H), show percentages of Enterobacterales infection isolates out of all bacterial
363 infection isolates with speciation results identified from blood and cerebrospinal fluid samples in target age
364 groups, in each setting, from 2010-2017 (Cambodia [n=197]; Kenya [n=910]; UK [n=3356]).
365

366 The highest relative abundances of resistance genes observed in each setting were for genes
367 associated with resistance to aminoglycosides, amphenicols, fluoroquinolones, tetracyclines and
368 macrolides (48.1%, 45.8% and 43.6% of the total counts in Cambodia, Kenya and the UK
369 respectively) (Fig 3A, left-hand panel). Relative abundance of resistance genes associated with these
370 five broad antibiotic classes differed between settings. For example, the relative abundance of
371 resistance genes for aminoglycosides in Cambodia (18.4%) was almost double that in Kenya (10.8%)
372 or UK (10.9%). The next highest relative abundance was of genes conferring resistance to penicillins
373 (Cambodia: 4.1%; Kenya: 4.7%; UK: 5.0%) and cephalosporins (Cambodia: 2.6%; Kenya: 2.3%; UK:
374 2.2%). Resistance gene counts for other antibiotic classes were <2% of the total gene counts in all
375 settings, including to carbapenems (Kenya [0.5%], Cambodia and UK [0.4%]). For single antibiotics
376 or antibiotic sub-classes (e.g. 1st generation cephalosporins), the highest relative abundances were
377 observed for erythromycin (Cambodia: 3.9%; Kenya: 4.2%; UK: 4.4%) and chloramphenicol
378 (Cambodia: 3.6%; Kenya: 3.5%; UK: 4.2%) in all settings (Fig 3A, right-hand panel). That for
379 resistance genes to antibiotics other than those listed was 76% (Cambodia), 76.5% (Kenya) and 76.1%
380 (UK) (data not shown). The relative abundance of resistance genes for all other single
381 antibiotics/antibiotic sub-classes was <2% in all settings, except for tigecycline (Cambodia: 2.8%;
382 Kenya and UK: 2.2%) and clindamycin (Kenya: 2.1%; UK: 2.6%). Resistance prevalence in
383 Enterobacterales isolates causing blood and cerebrospinal fluid infections is displayed in Fig.3B
384 (right-hand panel) for the 16 antibiotics with antibiotic susceptibility test data in ≥2 settings. For
385 comparison, this is shown alongside the relative abundance of resistance genes for the same
386 antibiotics in population pools (Fig.3B, left-hand panel).

387

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

388 **Fig 3. Relative abundance of resistance gene counts in population pools and percentage of**
389 **resistant Enterobacterales blood and cerebrospinal fluid infections in Cambodia, Kenya and**
390 **UK.**
391
392 Panels in Fig 3A show, for each setting, corrected resistance gene counts (CGCs) for a given antibiotic, antibiotic
393 class, or sub-class, divided by the total corrected AMR gene counts identified in the population pool. Relative
394 abundances were calculated using AMR_{ALL}, which considers corrected counts of genes and variants (CGC)
395 increasing the MIC or conferring clinically relevant resistance for a given antibiotic. Panels in Fig 3B show, for
396 each setting, the observed percentage of Enterobacterales resistant infections for 16 antibiotics with AST data in
397 ≥ 2 settings (right-hand side), and the relative abundance of CGCs for the same antibiotics in population pools,
398 based on AMR_{ALL} (left-hand side). Percentages are shown with 95% exact binomial confidence intervals in both
399 panels.
400

401 Pair-wise dissimilarities in resistomes from population pools, 30-sample-pools and individual sample
402 means (i.e. sum of CGC for the resistance gene types across all individually sequenced samples) were
403 calculated both within and across settings (Fig 4A and 4B), considering either the absolute CGC
404 values for each resistance gene type or their relative abundance based on the CGC values. Population
405 pools, 30-sample pools and individual sample means were less dissimilar and hence more closely
406 related within settings than across settings. In addition, within each setting, individual sample means
407 were more often less dissimilar to 30-sample pools than to population pools. In Cambodia 362 AMR
408 genes were identified in the 30-sample pool compared to 616 across all 30 individual samples. The
409 30-sample pool in Kenya comprised 339 genes compared to 499 across all individual samples.
410 Finally, in the UK 318 AMR genes were identified from the 30-sample pool compared to 422 across
411 all individual samples. However, when comparing individual samples and pools from the same setting
412 quantitatively, the average fraction of resistance genes for which the 30-sample pool estimate was
413 within the central interval of the empirical distribution inferred from individually sequenced samples
414 was 97% (Kenya: 98%; Cambodia: 97%; UK: 95%). In contrast, the average fraction was 86% across
415 comparisons between different settings (min-max: 80-92%). All 30-sample pool resistomes therefore
416 had substantially higher similarity to individual resistomes from the same setting relative to the
417 comparison with other settings.
418

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

419 **Fig 4. Pair-wise dissimilarities in the resistome of population pools, 30-sample-pools and**
420 **individual sample means within and across settings**
421
422 Using non-metric multidimensional scaling (NMDS) ordination-based method, Fig 4 shows pair-wise
423 dissimilarities of resistance gene counts from population pools (PP), 30-sample-pools (30S) and
424 individual sample means (SI) within and between settings, following mapping of sequences from
425 individual and pool metagenomes against CARD and a correction to remove resistance gene length
426 bias from counts. Dissimilarities are shown for the absolute corrected resistance gene counts (CGC;
427 left hand-side) and the relative abundance of resistance genes (right hand-side). Relative abundances
428 for genes in pools were calculated by dividing the CGC for each gene by the total CGC of all
429 resistance genes in the pool. Individual sample means were, for each resistance gene, the sum of CGC
430 across all individually sequenced samples. This, divided by the total CGC of all resistance genes
431 across all individually sequenced samples, was the relative abundance of each resistance gene based
432 on individual sample means.
433

434 The best taxonomy-adjusted RP metric - resulting in the highest point-wise out of sample prediction
435 accuracy and the greatest relative model weight - used the taxonomic parameter R_{tax} measuring
436 *Escherichia coli*, *Klebsiella pneumoniae*, Salmonella spp. and Enterobacter spp., and the abundance
437 of resistance genes increasing the MIC or conferring clinically relevant resistance (AMR_{ALL} version
438 of the R_{CGC} metric). This AMR_{ALL} model outperformed the other models, including a baseline model
439 without any metagenomics information, plus those models without taxonomic (R_{tax}) information
440 (Bayesian model averaging weights: Baseline [no R_{CGC} and no R_{tax}] = 0; R_{CGC} only [No R_{tax}] = 0; Best
441 model = 0.47]. Supplementary Data 4).

442

443 Model predictions were made for 16 antibiotics, which were those that had antibiotic susceptibility
444 test (AST) data for Enterobacterales isolates causing infection in at least two of the three settings
445 (Supplementary Data 5). Our best model accurately predicted the number of resistant infections in the
446 target populations for 100% of antibiotics with AST data in Kenya (12/12) and UK (14/14). In
447 Cambodia, the model accurately predicted the counts of resistant infections for 75% of antibiotics
448 (9/12). Compared to this, the baseline model did not correctly predict 50% of antibiotics across the
449 three settings (19/38). We computed the mean-squared errors of the mean model predictions relative
450 to the observations. The baseline model had an error of 468, whilst the final model (Fig 5) had an
451 error of 33.

452

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

453 **Fig 5. Bayesian model prediction of numbers of Enterobacterales invasive infections with**
454 **resistance to antibiotics with antibiotic susceptibility test (AST) results in ≥2 settings.**

455
456 Horizontal bars represent 95% highest density posterior interval and vertical lines represent means of the
457 predicted resistant sample counts based on the model using metagenomic data from population pools. Coloured
458 bars (yellow: Cambodia; blue: Kenya; brown: UK) are shown where clinical data on resistance (i.e. AST) was
459 available and grey bars where it was not. For grey bars the sample size was imputed. Red circles show the
460 number of blood and cerebrospinal fluid Enterobacterales infections that were found to be resistant to the
461 antibiotic listed in the y-axis. The number of isolates with AST results are also given in the y-axis. The red circle is
462 missing where no AST results were available. In cases where there is minimal uncertainty in the model estimate,
463 the red circle may overshadow the 95% credible interval bars (e.g. meropenem [Cambodia]; cefuroxime [Kenya]).
464 "Trimethoprim." is short for trimethoprim-sulfamethoxazole; "Cloramph" is short for chloramphenicol. NT = no
465 AST data available.
466

467 Bayesian model predictions expressed as percentages are shown in Supplementary Fig 2 for
468 antibiotics where AST results were available from > 100 invasive infection isolates. Above this
469 threshold, predicted percentage resistance was accurate for 100% of antibiotics (14/14 with >100
470 tested isolates).

471
472 ## Discussion

473
474 In this study we have demonstrated the feasibility of a novel, pragmatic approach to surveillance of
475 bacterial antimicrobial resistance of relevance to human infection, with a focus on Enterobacterales as
476 one of the major bacterial resistance threats[38,39]. Our results show that metagenomic analysis of pooled
477 faecal material (pooled at equimolar concentrations) is effective at predicting invasive infections
478 caused by Enterobacterales resistant to in-use antibiotics in a population, across a range of different
479 age groups and geographic settings. Our approach would enable intermittent, acceptable and relatively
480 non-invasive sampling of a small number of individuals within a population (e.g. 100-200), with the
481 advantage that a single centralised infrastructure (either in-country or internationally) could undertake
482 the metagenomic sequencing and analysis. This can be done independently of development of a
483 network of classical microbiological laboratories in multiple settings, which can be resource-intensive
484 in terms of capital and running costs, and is not feasible in the short-term, especially in LMICs, which
485 frequently have the highest AMR burden.

486

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

487 Based solely on pool size and sequencing depth (50-55Gbp/pool), we developed predictive metrics
488 (RP) without the need for costly and labour-intensive multiplexing of samples (i.e. individually
489 identifying samples in the pool by means of barcoded sequences) or selective sequencing approaches
490 based on enrichment for predefined panels of resistance genes. Unlike other AMR gene profiling
491 approaches our bioinformatics pipeline (ResPipe) incorporates the capacity to identify both specific
492 AMR gene variants (e.g. such as $bla_{\text{CTX-M-33}}$ versus $bla_{\text{CTX-M-63}}$), as well as being able to aggregate by
493 gene family. This is especially important for the prediction of phenotypes, as genes that differ by only
494 single nucleotides/amino acids can have distinct phenotypic spectra. Pooled metagenomes/resistomes
495 were also found to be an accurate, non-biased representation of the individual sample
496 metagenomes/resistomes. Population pools comprising rectal swabs with as little as ≥1ng/ul
497 DNA/sample were found to be sufficient to derive RP metrics with predictive value; this is useful in
498 terms of optimizing the sample processing workflows. Finally, in producing relatively deeply
499 sequenced (50-55Gbp/metagenome) and complete (i.e. not restricted to 16S) metagenomes on 90
500 individuals, we have also made a significant contribution to the human microbiomics data repository,
501 freely available for other researchers to use for study.

502

503 The limitations of our approach were most obvious for the neonatal group from Cambodia, where
504 predicted resistance matched the observed resistance in invasive isolates for 75% of antibiotics
505 compared to 100% of antibiotics in Kenya and the UK. One explanation for this might be that the
506 population pool for this group was found to have included 19 longitudinal samples (12% of all
507 samples in the pool) collected from individuals during their hospital inpatient stay, potentially biasing
508 the metagenomics profile of population pools and infection metadata designed to reflect community
509 (i.e. non-hospital) profiles. Rapid changes in the neonatal resistome occur following exposure to the
510 hospital environment[40]. Analyses of neonatal metagenomes have shown that these are predisposed to
511 rapid flux, and in hospital typically reflect the environmental hospital "microbiome"[41]. Cambodia was
512 also the only setting where the age group considered for metagenomics analysis (i.e. neonates), did
513 not correspond exactly with the available infection metadata analysed (i.e. infants up to 90 days of

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

514 age), which may also have influenced the accuracy of our predictive approach. Our analysis was also
515 limited by the scarce antibiotic susceptibility test (AST) results available for invasive infection
516 isolates, particularly in Cambodia, where the maximum number of isolates with AST results for any
517 given antibiotic was 65, compared to 324 in Kenya and 912 in UK. The smaller number of isolates
518 from Cambodia meant that the model fit contained less information to accurately predict resistance in
519 this setting. Moreover, AST results were only available for a limited number of antibiotics across all
520 three settings, and ideally AST approaches used for comparison would have been standardised across
521 the settings. Finally, our analyses are heavily dependent on the robustness of the reference gene
522 database, and the accuracy of genotypic-phenotypic correlations catalogued therein. In general,
523 however, we would expect this knowledge base to become increasingly robust, thus strengthening our
524 predictions. This may explain why in this study, a model that considers all gene variants with
525 experimental evidence of increasing the minimum inhibitory concentration (MIC), outperformed a
526 model considering only genes known to confer clinically relevant resistance.

527

528 Further studies to validate our promising proof-of-principle observations in additional settings across
529 age categories, especially the neonatal group, are warranted. There is potential to extend the approach
530 to consider other priority bacterial groups and different colonisation samples. For example, pools
531 could be extended to include samples from nasopharyngeal sites, where other potential pathogens
532 predominate (e.g. *Streptococcus* spp., *Staphylococcus* spp.). To develop the most rapid, convenient,
533 simple and inexpensive method possible, future studies should also consider further simplifications to
534 the method such as whether the same accurate predictions can be generated by pooling all samples
535 prior to DNA extraction and then performing the extraction only once. Further work should also test
536 the resolution of the approach to characterise and track local/sub-national variation in AMR
537 prevalence, or in community versus healthcare-associated contexts. A mathematical framework for
538 minimum-cost implementation of pooled-sample metagenomics-based surveys to quantify the burden
539 of resistance in new settings without prior microbiology or AST data would also be of benefit, and

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

540 could be greatly informed by the data we have generated, which can contribute to simulation work
541 addressing pools sizes, pool numbers per region, and sequencing depth.
542
543 We conclude that surveillance based on population colonisation metagenomics and taxonomy-
544 adjusted AMR metrics presented here are in principle a valuable public health opportunity, and may
545 represent an alternative or bridging measure to the implementation of local and regional laboratory-
546 based infrastructures focussed on culturing isolates from clinical specimens, especially in resource-
547 limited settings. This novel approach could be used to overcome the current paucity of quality AMR
548 surveillance data and inform setting-tailored rationalization of/or access to antibiotics, context-
549 appropriate treatment guidelines, organized measures to prevent AMR and ultimately public-health
550 decision in conjunction with relevant stakeholders, especially in LMICs.
551
552 **Ethics Declarations**
553 This research was conducted with approval from the Oxford Tropical Research Ethics Committee
554 (OxTREC Reference: 5126-16) following local ethics clearance for inclusion of Cambodia and Kenya
555 sample collections, and approval of a substantial amendment to 14/LO/2085 by the National Research
556 Ethics Service (NRES London – Camberwell St Giles), for inclusion of the London sample collection
557 in this study.
558
559 The authors declare no competing interests.
560
561 **Data and Code Availability**
562 The raw sequence data reported in this study have been deposited in the European Nucleotide Archive
563 under accession number PRJEB34871. The code to extract CARD data, that required to generate the
564 final datasets and analyses, plus any required input files, are available from the ResPipe GitLab
565 repository (https://gitlab.com/hsgweon/ResPipe); ResPipe output data can be found at the ResPipe
566 Gitlab subdirectory (https://gitlab.com/hsgweon/ResPipe/tree/master/data).
567

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

568 **Author Contributions**

569 This work was first conceived by O.T.A., with support from N.S. and B.S.C.; O.T.A, N.S., B.S.C.,
570 R.N. and H.S.G. designed the study. K.C., J.W., O.T.A. and N.S. developed and validated modified
571 DNA extraction protocols for this study. K.C., J.W. and R.B. conducted or facilitated most of the pre-
572 sample-pooling laboratory work. S.L. designed the methods and provided technical guidance for
573 sample pooling and sequencing and conducted the sequencing work. J.A.B., J.D.E., P.T. and R.B.
574 facilitated the collation and transfer of samples and data from participant settings. They also provided
575 technical support for clinical and microbiology study procedures and for the development of context-
576 appropriate standard operating procedures. N.S., A.S.W., T.E.P., D.W.C. and B.S.C. provided support
577 and guidance for all technical aspects of the study (including for bioinformatics and data analyses)
578 and contributed to the revision of study outputs. T.N. contributed to the mining, standardisation and
579 analysis of infection metadata from each setting. H.S.G. conducted the bioinformatics work, designed
580 the methods for corrected gene counts and extracted the data from CARD. J.S. provided the
581 computing support for the study. O.T.A. conducted mining, linkage and visualisation of study data.
582 R.N. conducted the validation and Bayesian analyses and B.S.C. contributed to revision of these
583 methods. O.T.A, N.S., R.N. and H.S.G. produced the first manuscript draft. All authors contributed
584 significantly to the iterative review of the draft.

585

586 **Acknowledgements**

587 We are grateful to Professor Mike English (Centre for Tropical Medicine and Global Health,
588 University of Oxford, UK), for facilitating the initiation of this work by bringing together some of the
589 groups involved in the research. We are also grateful to Dr Alessandra Natale, for supporting the
590 laboratory work for this study at the CIDR (Centre for Clinical Infection and Diagnostics Research,
591 Department of Infectious Diseases, King's College London, UK). The study was funded by Bill &
592 Melinda Gates Foundation (grant agreement OPP1160974) and was sponsored by University of
593 Oxford.

594

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

595 **References**

596 1. World Health Organisation. *Global action plan on Antimicrobial Resistance*.
597 http://www.who.int/antimicrobial-resistance/global-action-plan/en/ (2016).

598 2. Collignon, P., Beggs, J. J., Walsh, T. R., Gandra, S. & Laxminarayan, R. Anthropological and
599 socioeconomic factors contributing to global antimicrobial resistance: a univariate and
600 multivariable analysis. *Lancet Planet. Heal.* **2**, e398–e405 (2018).

601 3. World Health Organisation. *Global Antimicrobial Resistance Surveillance System (GLASS)*.
602 http://www.who.int/drugresistance/en/ (2015).

603 4. Wertheim, H. F. L. *et al.* Risk and outcome of nosocomial Staphylococcus aureus bacteraemia
604 in nasal carriers versus non-carriers. *Lancet* **364**, 703–705 (2004).

605 5. Tischendorf, J., de Avila, R. A. & Safdar, N. Risk of infection following colonization with
606 carbapenem-resistant Enterobactericeae: A systematic review. *Am. J. Infect. Control* **44**, 539–
607 543 (2016).

608 6. Forslund, K. *et al.* Country-specific antibiotic use practices impact the human gut resistome.
609 *Genome Res.* **23**, 1163–9 (2013).

610 7. Korpela, K. *et al.* Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-
611 school children. *Nat. Commun.* **7**, 10410 (2016).

612 8. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics
613 analyses of urban sewage. *Nat. Commun.* **10**, (2019).

614 9. Forslund, K., Sunagawa, S., Coelho, L. P. & Bork, P. Metagenomic insights into the human
615 gut resistome and the forces that shape it. *Bioessays* **36**, 316–29 (2014).

616 10. Standing, J. F. *et al.* Dosing of Ceftriaxone and Metronidazole for Children With Severe Acute
617 Malnutrition. *Clin. Pharmacol. Ther.* **104**, 1165–1174 (2018).

618 11. Turner, P. *et al.* High prevalence of antimicrobial-resistant gram-negative colonization in
619 hospitalized cambodian infants. *Pediatr. Infect. Dis. J.* **35**, 856–861 (2016).

620 12. Otter, J. A. *et al.* Universal hospital admission screening for carbapenemase-producing
621 organisms in a low-prevalence setting. *J. Antimicrob. Chemother.* **71**, (2016).

622 13. Dyakova, E. *et al. Efficacy and acceptability of rectal and perineal sampling for identifying*
623 *gastrointestinal colonisation with ESBL-Enterobacteriaceae*. *Clinical Microbiology and*
624 *Infection* http://www.sciencedirect.com/science/article/pii/S1198743X17301167 (2017)
625 doi:10.1016/j.cmi.2017.02.019.

626 14. Otter, J. A. *et al.* Individual- and community-level risk factors for ESBL Enterobacteriaceae
627 colonization identified by universal admission screening in London. *Clin. Microbiol. Infect.* **0**,
628 (2019).

629 15. Budding, A. E. *et al.* Rectal swabs for analysis of the intestinal microbiota. *PLoS One* **9**,
630 e101344 (2014).

631 16. Baumann-Dudenhoeffer, A. M., D'Souza, A. W., Tarr, P. I., Warner, B. B. & Dantas, G. Infant
632 diet and maternal gestational weight gain predict early metabolic maturation of gut
633 microbiomes. *Nature Medicine* vol. 24 1822–1829 (2018).

634 17. Bassis, C. M. *et al.* Comparison of stool versus rectal swab samples and storage conditions on
635 bacterial community profiles. *BMC Microbiol.* **17**, (2017).

636 18. Chheng, K. *et al.* A Prospective Study of the Causes of Febrile Illness Requiring
637 Hospitalization in Children in Cambodia. *PLoS One* **8**, (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

638 19. Fox-Lewis, A. *et al.* Antimicrobial resistance in invasive bacterial infections in hospitalized
639 children, Cambodia, 2007–2016. *Emerg. Infect. Dis.* **24**, 841–851 (2018).

640 20. Scott, J. A. G. *et al.* Profile: The Kilifi health and demographic surveillance system (KHDSS).
641 *Int. J. Epidemiol.* **41**, 650–657 (2012).

642 21. Tosas Auguet, O. *et al.* Frequent Undetected Ward-Based Methicillin-Resistant
643 Staphylococcus aureus Transmission Linked to Patient Sharing Between Hospitals. *Clin.*
644 *Infect. Dis.* **66**, 840–848 (2018).

645 22. British Society for Antimicrobial Chemotherapy. http://www.bsac.org.uk/.

646 23. Clinical and Laboratory Standards Institute (CLSI). *Performance Standards for Antimicrobial*
647 *Susceptibility Testing. CLSI document M100-S27.* (Wayne, PA: Clinical and Laboratory
648 Standards Institute, 2017).

649 24. Satinsky, B. M., Gifford, S. M., Crump, B. C. & Moran, M. A. Chapter Twelve – Use of
650 Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. in *Methods*
651 *in Enzymology* vol. 531 237–250 (2013).

652 25. Lamble, S. *et al.* Improved workflows for high throughput library preparation using the
653 transposome-based nextera system. *BMC Biotechnol.* **13**, (2013).

654 26. Gweon, H. S. *et al.* The impact of sequencing depth on the inferred taxonomic composition
655 and AMR gene content of metagenomic samples. *Environ. Microbiome* **14**, 7 (2019).

656 27. Babraham Bioinformatics - Trim Galore!
657 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

658 28. Wood, D. E. *et al.* Kraken: ultrafast metagenomic sequence classification using exact
659 alignments. *Genome Biol.* **15**, R46 (2014).

660 29. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance
661 in metagenomics data. *PeerJ* **2017**, (2017).

662 30. Jia, B. *et al.* CARD 2017: Expansion and model-centric curation of the comprehensive
663 antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).

664 31. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive
665 antibiotic resistance database. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz935.

666 32. BBMap: A Fast, Accurate, Splice-Aware Aligner (Conference) | OSTI.GOV.
667 https://www.osti.gov/biblio/1241166.

668 33. Martínez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene? Ranking risk in
669 resistomes. *Nat. Rev. Microbiol.* **13**, 116–23 (2015).

670 34. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, (2017).

671 35. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out
672 cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).

673 36. Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. Using Stacking to Average Bayesian
674 Predictive Distributions (with Discussion). *Bayesian Anal.* **13**, 917–1007 (2018).

675 37. R Core Team. R: A language and environment for statistical computing. (2018).

676 38. World Health Organisation. *Prioritization of pathogens to guide discovery, research and*
677 *development of new antibiotics for drug-resistant bacterial infections, including tuberculosis*.
678 https://www.who.int/medicines/areas/rational_use/prioritization-of-pathogens/en/ (2017).

679 39. Tacconelli, E. *et al.* Discovery, research, and development of new antibiotics: the WHO

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

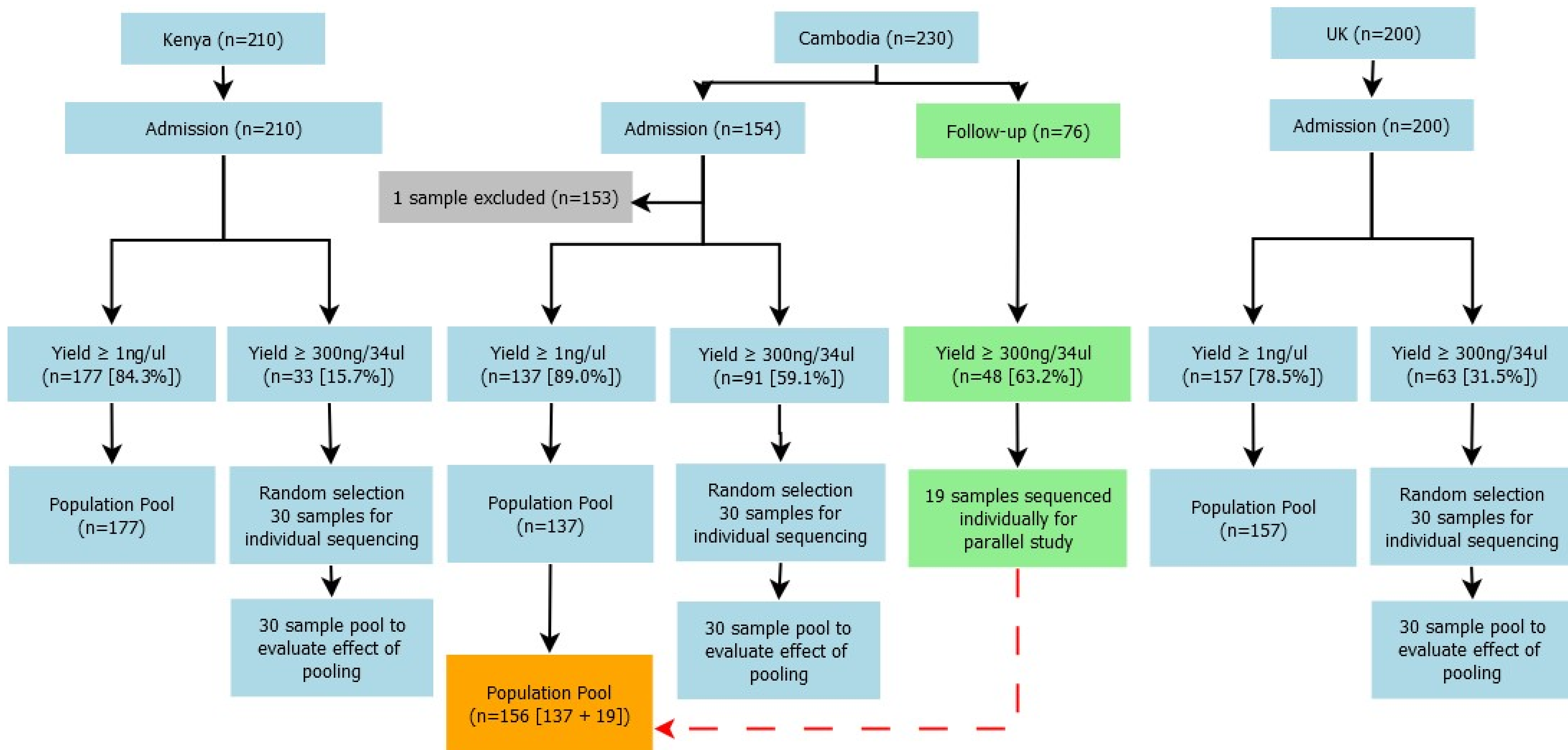
680 priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327
681 (2018).

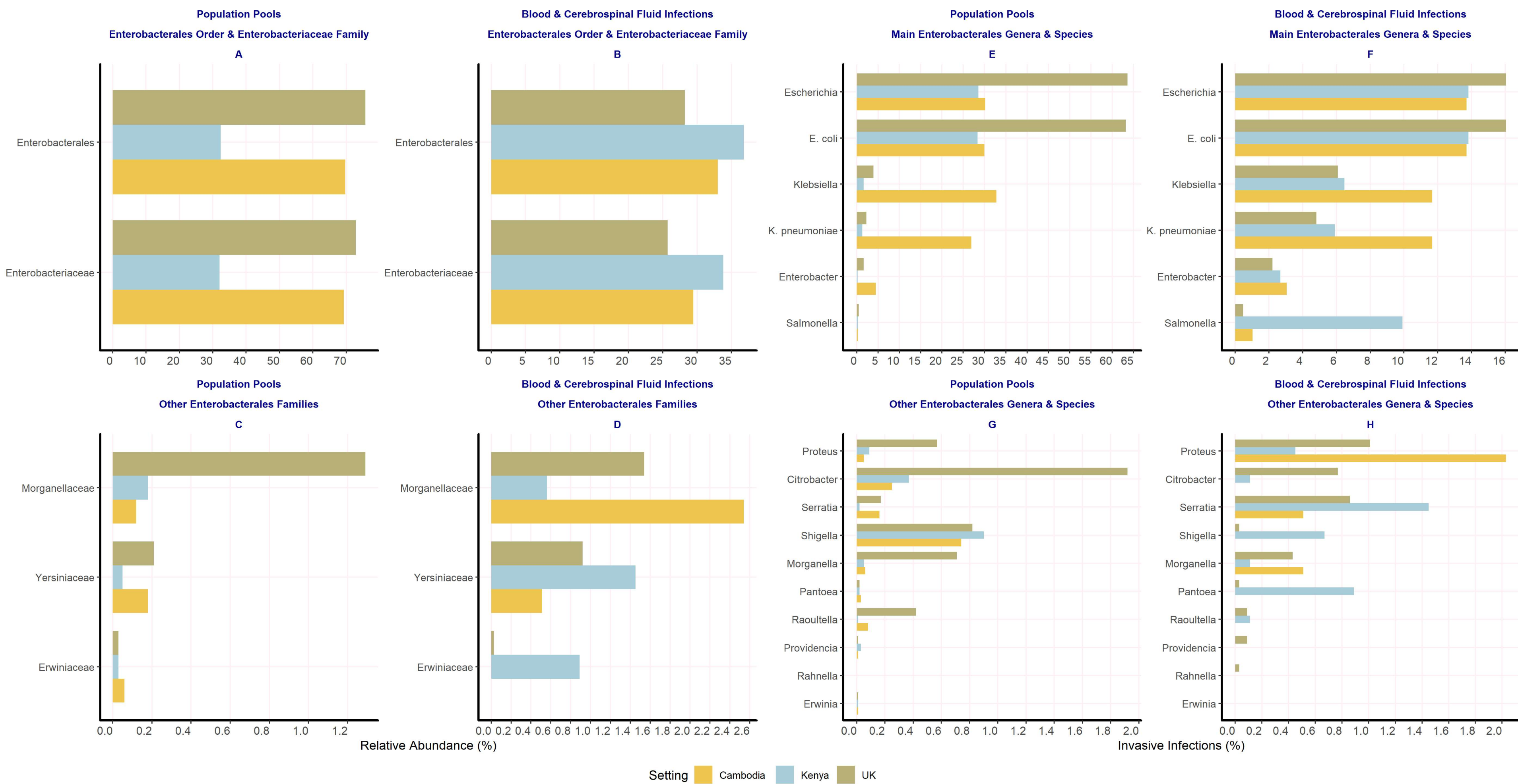
682 40. Kagia, N. *et al.* Carriage and Acquisition of Extended-spectrum β-Lactamase-producing
683 Enterobacterales among Neonates Admitted to Hospital in Kilifi, Kenya. *Clin. Infect. Dis.* **69**,
684 751–759 (2019).

685 41. Brooks, B. *et al.* Strain-resolved analysis of hospital rooms and infants reveals overlap
686 between the human and room microbiome. *Nat. Commun.* **8**, (2017).

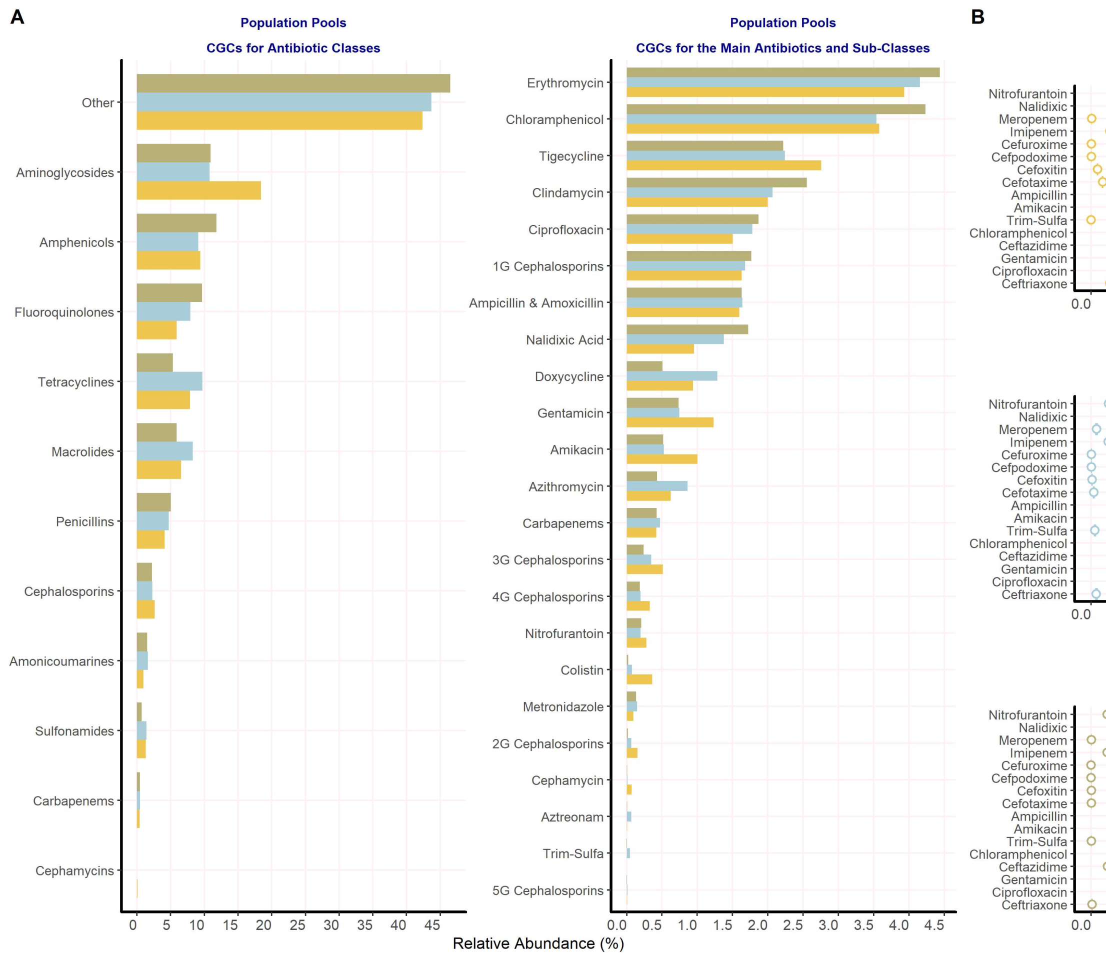
687

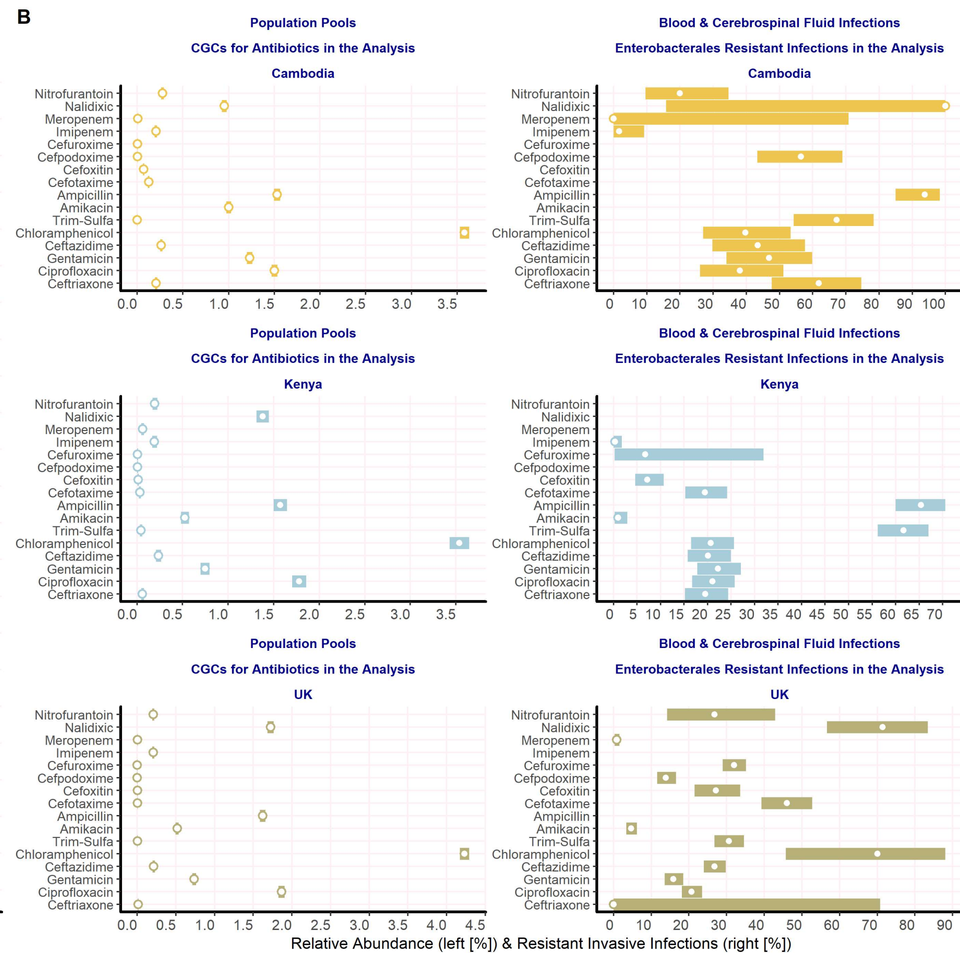
688

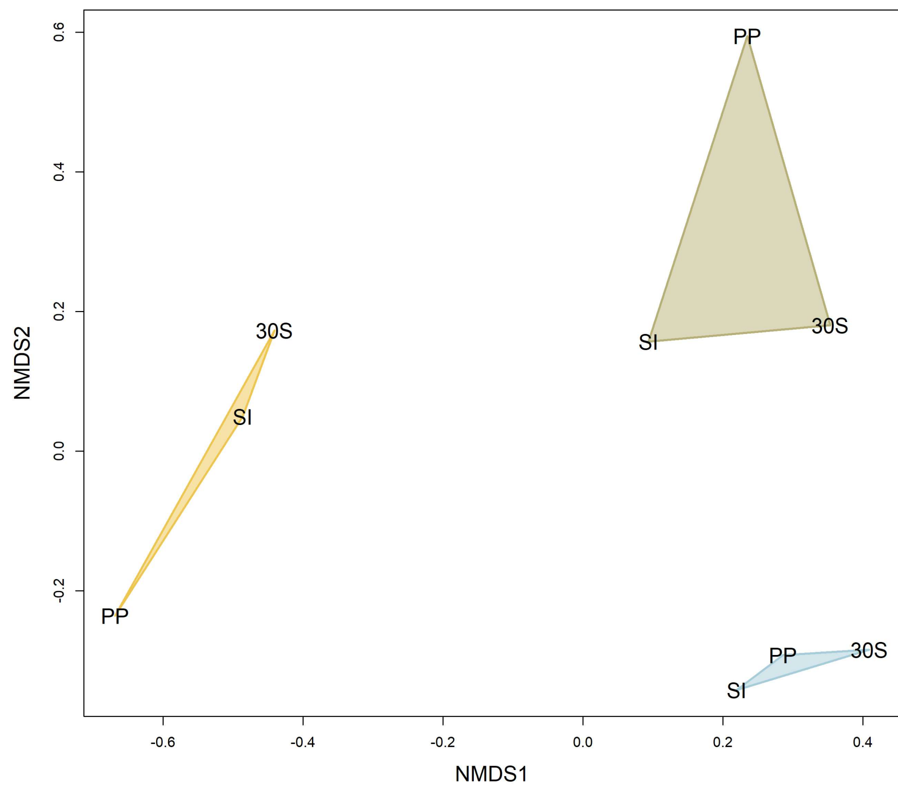

Kenya (n=210)
Admission (n=210)
Yield ≥ 1ng/ul (n=177 [84.3%])
Yield ≥ 300ng/34ul (n=33 [15.7%])
Population Pool (n=177)
Random selection 30 samples for individual sequencing
30 sample pool to evaluate effect of pooling
Cambodia (n=230)
Admission (n=154)
Follow-up (n=76)
1 sample excluded (n=153)
Yield ≥ 1ng/ul (n=137 [89.0%])
Yield ≥ 300ng/34ul (n=91 [59.1%])
Yield ≥ 300ng/34ul (n=48 [63.2%])
Population Pool (n=137)
Random selection 30 samples for individual sequencing
19 samples sequenced individually for parallel study
30 sample pool to evaluate effect of pooling
Population Pool (n=156 [137 + 19])
UK (n=200)
Admission (n=200)
Yield ≥ 1ng/ul (n=157 [78.5%])
Yield ≥ 300ng/34ul (n=63 [31.5%])
Population Pool (n=157)
Random selection 30 samples for individual sequencing
30 sample pool to evaluate effect of pooling

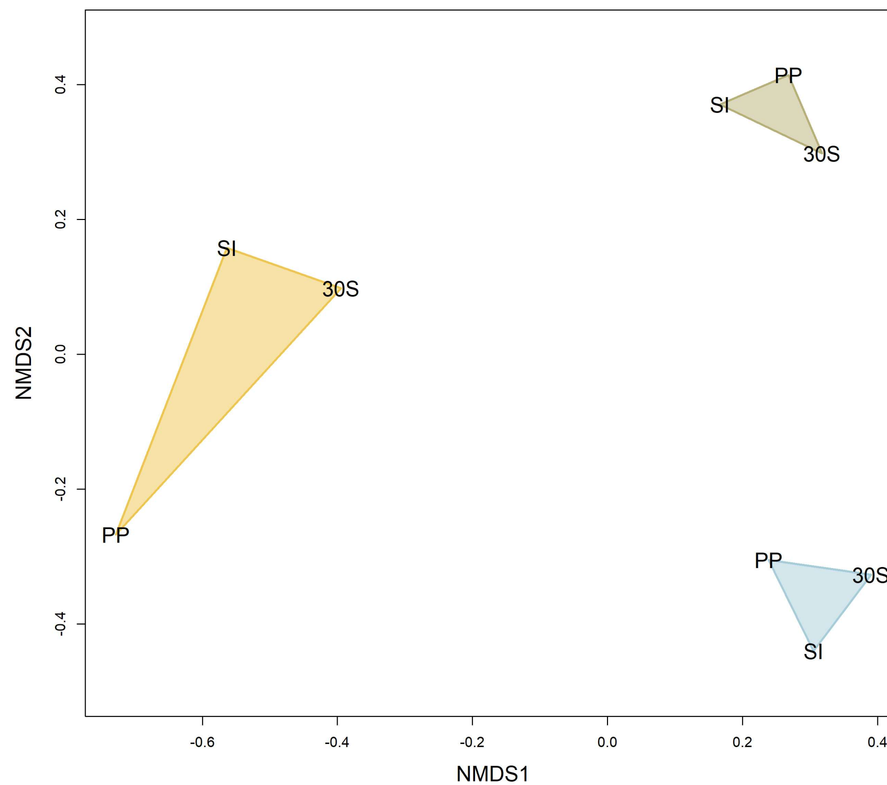


Population Pools
Enterobacterales Order & Enterobacteriaceae Family
A
Blood & Cerebrospinal Fluid Infections
Enterobacterales Order & Enterobacteriaceae Family
B
Population Pools
Main Enterobacterales Genera & Species
E
Blood & Cerebrospinal Fluid Infections
Main Enterobacterales Genera & Species
F
Enterobacterales
Enterobacteriaceae
Escherichia
E. coli
Klebsiella
K. pneumoniae
Enterobacter
Salmonella
Population Pools
Other Enterobacterales Families
C
Blood & Cerebrospinal Fluid Infections
Other Enterobacterales Families
D
Population Pools
Other Enterobacterales Genera & Species
G
Blood & Cerebrospinal Fluid Infections
Other Enterobacterales Genera & Species
H
Morganellaceae
Yersiniaceae
Erwiniaceae
Proteus
Citrobacter
Serratia
Shigella
Morganella
Pantoea
Raoultella
Providencia
Rahnella
Erwinia
Relative Abundance (%)
Invasive Infections (%)
Setting
Cambodia
Kenya
UK


A
Population Pools
CGCs for Antibiotic Classes
Other
Aminoglycosides
Amphenicols
Fluoroquinolones
Tetracyclines
Macrolides
Penicillins
Cephalosporins
Amonicoumarines
Sulfonamides
Carbapenems
Cephamycins
Population Pools
CGCs for the Main Antibiotics and Sub-Classes
Erythromycin
Chloramphenicol
Tigecycline
Clindamycin
Ciprofloxacin
1G Cephalosporins
Ampicillin & Amoxicillin
Nalidixic Acid
Doxycycline
Gentamicin
Amikacin
Azithromycin
Carbapenems
3G Cephalosporins
4G Cephalosporins
Nitrofurantoin
Colistin
Metronidazole
2G Cephalosporins
Cephamycin
Aztreonam
Trim-Sulfa
5G Cephalosporins
Relative Abundance (%)

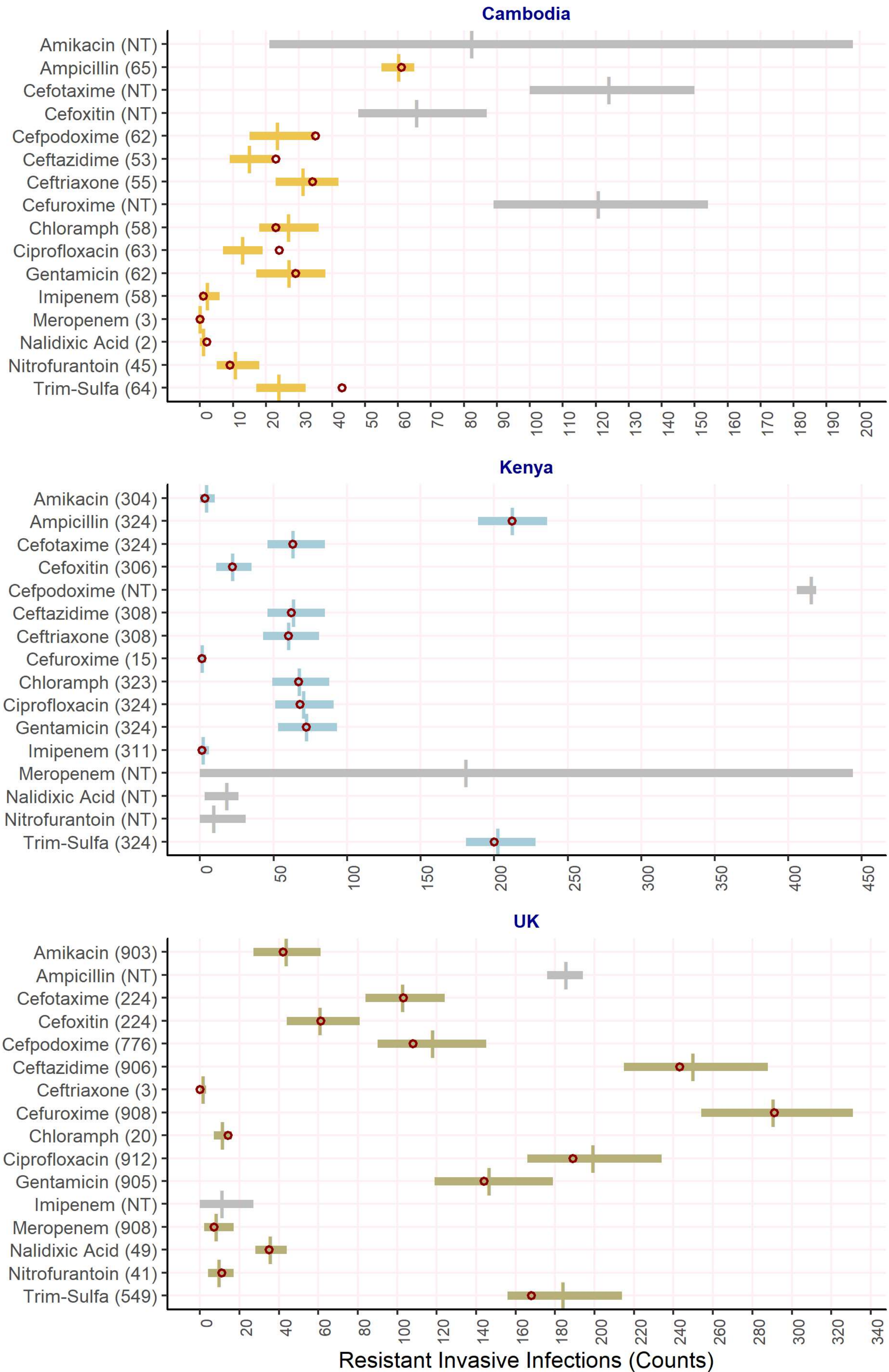

B
Population Pools
CGCs for Antibiotics in the Analysis
Cambodia
Kenya
UK
Blood & Cerebrospinal Fluid Infections
Enterobacterales Resistant Infections in the Analysis
Nitrofurantoin
Nalidixic
Meropenem
Imipenem
Cefuroxime
Cefpodoxime
Cefoxitin
Cefotaxime
Ampicillin
Amikacin
Trim-Sulfa
Chloramphenicol
Ceftazidime
Gentamicin
Ciprofloxacin
Ceftriaxone
Relative Abundance (left [%]) & Resistant Invasive Infections (right [%])
Setting
Cambodia
Kenya
UK

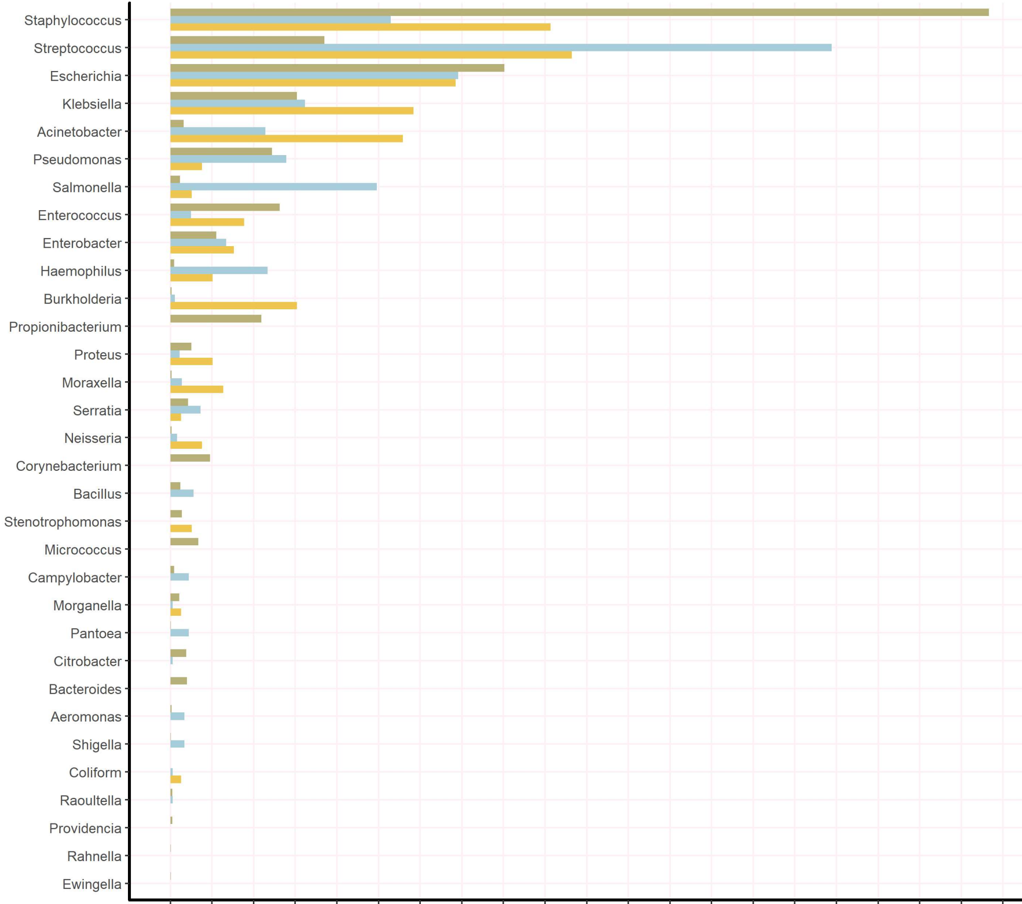
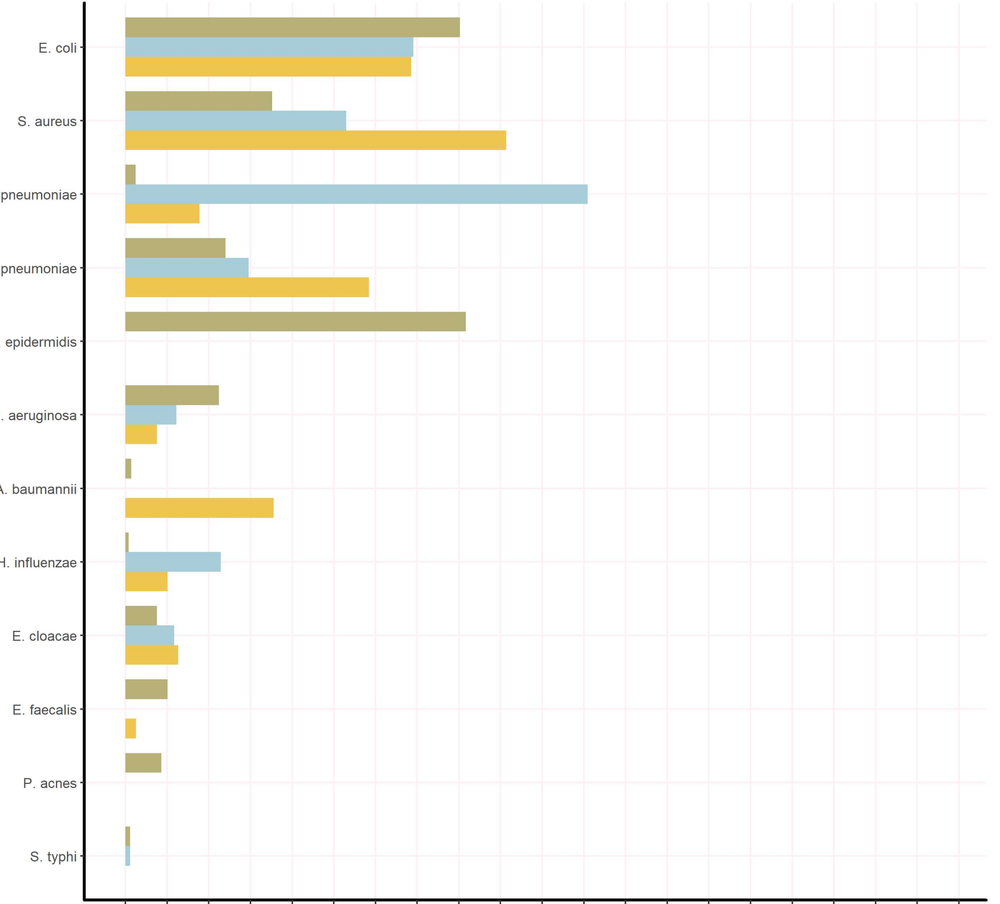




Cambodia Kenya UK

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.




Blood & Cerebrospinal Fluid Infections
All Genera
Staphylococcus
Streptococcus
Escherichia
Klebsiella
Acinetobacter
Pseudomonas
Salmonella
Enterococcus
Enterobacter
Haemophilus
Burkholderia
Propionibacterium
Proteus
Moraxella
Serratia
Neisseria
Corynebacterium
Bacillus
Stenotrophomonas
Micrococcus
Campylobacter
Morganella
Pantoea
Citrobacter
Bacteroides
Aeromonas
Shigella
Coliform
Raoultella
Providencia
Rahnella
Ewingella
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40

Blood & Cerebrospinal Fluid Infections
All Species
E. coli
S. aureus
S. pneumoniae
K. pneumoniae
S. epidermidis
P. aeruginosa
A. baumannii
H. influenzae
E. cloacae
E. faecalis
P. acnes
S. typhi
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Invasive Infections (%)
Setting
Cambodia
Kenya
UK

bioRxiv preprint doi: https://doi.org/10.1101/2020.02.10.941930; this version posted February 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.