

# Personal clinical history predicts antibiotic resistance of urinary tract infections

Idan Yelin<sup>ID 1</sup>, Olga Snitser<sup>1</sup>, Gal Novich<sup>2</sup>, Rachel Katz<sup>3</sup>, Ofir Tal<sup>4</sup>, Miriam Parizade<sup>5</sup>, Gabriel Chodick<sup>ID 3,6</sup>, Gideon Koren<sup>3,6</sup>, Varda Shalev<sup>3,6</sup> and Roy Kishony<sup>ID 1,2,4\*</sup>

**Antibiotic resistance is prevalent among the bacterial pathogens causing urinary tract infections. However, antimicrobial treatment is often prescribed 'empirically', in the absence of antibiotic susceptibility testing, risking mismatched and therefore ineffective treatment. Here, linking a 10-year longitudinal data set of over 700,000 community-acquired urinary tract infections with over 5,000,000 individually resolved records of antibiotic purchases, we identify strong associations of antibiotic resistance with the demographics, records of past urine cultures and history of drug purchases of the patients. When combined together, these associations allow for machine-learning-based personalized drug-specific predictions of antibiotic resistance, thereby enabling drug-prescribing algorithms that match an antibiotic treatment recommendation to the expected resistance of each sample. Applying these algorithms retrospectively, over a 1-year test period, we find that they greatly reduce the risk of mismatched treatment compared with the current standard of care. The clinical application of such algorithms may help improve the effectiveness of antimicrobial treatments.**

The resistance of bacterial pathogens to commonly used antibiotics is a growing public health concern, threatening the efficacy of antibiotic drugs<sup>1,2</sup>. The use of antibiotics benefits resistant strains, exacerbating the problem over time<sup>3–7</sup>. At the single-patient level, the efficacy of antimicrobial treatment is critically dependent on correctly matching antibiotic choice to the specific susceptibilities of the pathogen<sup>8–10</sup>. Ideally, correct prescription should be based on direct measurement of the antibiotic susceptibilities of the infecting pathogen. In practice though, to provide rapid clinical intervention, drugs are often prescribed empirically in the absence of culture susceptibility measurements, risking incorrect and therefore ineffective treatment.

This problem is of particular importance in urinary tract infections (UTIs), one of the most frequent community-acquired infections worldwide, for which the common practice of empirical treatment is jeopardized by the substantial frequency of resistant infections. UTIs are among the most common bacterial infections, with over 150 million annual cases globally<sup>11</sup>. One in three women will have at least one symptomatic UTI by the age of 24 and more than half will be affected during their lifetime<sup>12</sup>. Treatment of these infections accounts for about 8% of non-hospital usage of antibiotics, often as part of empirical prescription<sup>13–15</sup>. The common etiological agents of UTIs are diverse, including *Escherichia coli*, *Klebsiella pneumoniae* and *Proteus mirabilis*, as well as Gram-positive bacteria such as *Enterococcus faecalis*<sup>16–21</sup>. These pathogens are often resistant to several antibiotics, with resistance rates of infections exceeding 20% for commonly used drugs<sup>17,20,22</sup>, emphasizing the challenge of empirically prescribing the specific antibiotics to which the infecting pathogen is susceptible<sup>23</sup>.

The risk of an infection being resistant to different antibiotics is associated with patient demographics and comorbidities. Known demographic factors associated with resistance include older age<sup>24</sup>, gender<sup>25</sup>, ethnicity<sup>26–29</sup>, residence in a retirement home<sup>25</sup> and travel

to developing countries<sup>28</sup>. Known comorbidities associated with resistance include the presence of a urinary catheter<sup>21,25,30</sup>, immunodeficiency<sup>25</sup> and diabetes<sup>25</sup>. Notably, most of these associations were identified based on small patient cohorts, typically with high frequencies of antibiotic-resistant infections, such as retirement homes, rehabilitation centers or hospitals.

Beyond the patient's demographics and comorbidities, antibiotic resistance has also been associated with the patient's past clinical history, including recurrent UTIs, hospitalizations and resistance of previous infections. Risk of resistance to specific drugs have been shown to increase for patients with recurrent UTIs<sup>25,29,31</sup> and past hospitalizations<sup>25,32</sup>. Studies have further shown that resistance of previous infections can be used to predict resistance in future infections<sup>33,34</sup>; however, the time extent of these associations is not well resolved and it is also unclear whether and how these associations vary across resistances to different antibiotics.

The availability of antibiotic purchase data reveals patterns of antibiotic use<sup>15,35</sup> and shows that the risk of resistance increases with short-term prior use of antibiotics<sup>5,24,25,32,36–38</sup>. Recent large-scale studies showed that, across geography, resistance levels can be correlated with past drug consumption<sup>20,39</sup>. Resistance to fluoroquinolones was correlated with past consumption volumes of these same drugs<sup>20</sup>, while resistance to trimethoprim-sulfa was correlated with the volume of consumption of the same drug (cognate) as well as of other drugs of different pharmaceutical classes (non-cognate)<sup>20</sup>. Such associations of usage of a given antibiotic with future resistance to other antibiotics can appear indirectly through co-occurrence among resistance mechanisms (for example, if resistance to drug X and resistance to drug Y are correlated, then direct selection by drug X to X-resistance may result in an association of drug X with resistance to drug Y). Resolving direct and indirect selection for resistance has been challenging in the absence of resistance co-occurrence data. Negative associations, where drug use is

<sup>1</sup>Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel. <sup>2</sup>Department of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel. <sup>3</sup>Maccabitech, Maccabi Healthcare Services, Tel-Aviv, Israel. <sup>4</sup>Lorry I. Lokey Interdisciplinary Center for Life Sciences & Engineering, Technion-Israel Institute of Technology, Haifa, Israel. <sup>5</sup>Maccabi Healthcare Services, National Laboratory, Rechovot, Israel. <sup>6</sup>Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel. \*e-mail: [rkishony@technion.ac.il](mailto:rkishony@technion.ac.il)

anti-correlated with resistance, have also been observed, but it has been difficult to discern the direction of causality<sup>20,40</sup>. Finally, the time extent of these positive and negative associations of resistance with prior antibiotic usage is not well resolved.

Here, we present an analysis of a large population of patients with UTIs to unravel predictive features of antibiotic resistance and test how these features can be combined to recommend optimal drugs for empirical treatment. We analyze a patient-level longitudinal data set of community and retirement home-acquired UTI cultures collected by Maccabi Healthcare Services (MHS), Israel's second largest health maintenance organization, serving a diverse population of ~2 million patients. Analyzing demographic factors, we find strong drug-specific associations with resistance. Then, comparing resistance data of multiple infections from the same patient, we unravel a decaying long-term memory-like correlation of resistance over time. We also combine these culture records with patient-linked records of antibiotic use to quantify the extent and time of direct and indirect correlations of antibiotic use with resistance at the single-patient level. Finally, combining these demographic and historical factors for personalized predictions of resistance, we develop machine-learning models that we demonstrate can substantially improve upon physician-prescribed empirical antibiotic treatment.

## Results

We retrieved data of all positive urine cultures of MHS patients for the 10-year period between 1 July 2007 and 30 June 2017, as well as patient demographics and records of antibiotic purchases for these patients (Methods). Among all ~2 million MHS patients, there were 711,099 recorded positive urine samples from 315,047 patients in total. For each positive sample, one or more bacterial species were isolated and characterized. The data set included species-level identification of these isolates as well as resistance profiles measured by VITEK 2, reinterpreted in accordance with The Clinical and Laboratory Standards Institute (CLSI) guidelines (sensitive, intermediate and resistant). Because a multi-species infection can be treated by a given drug only if none of the isolates is resistant to it, we define for each antibiotic and each sample the 'sample resistance': the maximal resistance across all isolates from the same sample (96.4% of samples were identified as single species and their resistance profile is simply defined as the resistance profile of their single isolates). All of MHS's country-wide clinical tests are performed centrally (Methods), allowing reliable comparison across patients and time. In our analysis, we focus on resistance to the six drugs that were most commonly prescribed as part of empirical treatment of these infections (identified as the drugs commonly given on the same day samples were sent for culture; Table 1 and Supplementary Table 1; Methods). Resistance measurements for these antibiotics were carried out routinely over the entire 10-year period (except for cephalexin for which measurements are available only since 2014; Extended Data Fig. 1).

Three species, *E. coli*, *K. pneumoniae* and *P. mirabilis*, account for 85% of isolates (70%, 10% and 5%, respectively; Fig. 1a). These pathogens varied in their resistance profiles (Fig. 1b). Notably, for all six antibiotics, the chance of resistant infection is substantial, indicating that antibiotic treatment efficacy could often be undermined. These population-level frequencies of resistance were fairly static over time (for example, trimethoprim-sulfa or nitrofurantoin) with only mild changes observed in certain antibiotics and specific species (Fig. 1c and Extended Data Fig. 2). The diversity of pathogens and resistance patterns underscores that antibiotic prescriptions must be tailored to match the resistance profile of the infection<sup>41</sup>, motivating the development of methods to better predict resistance<sup>23</sup>.

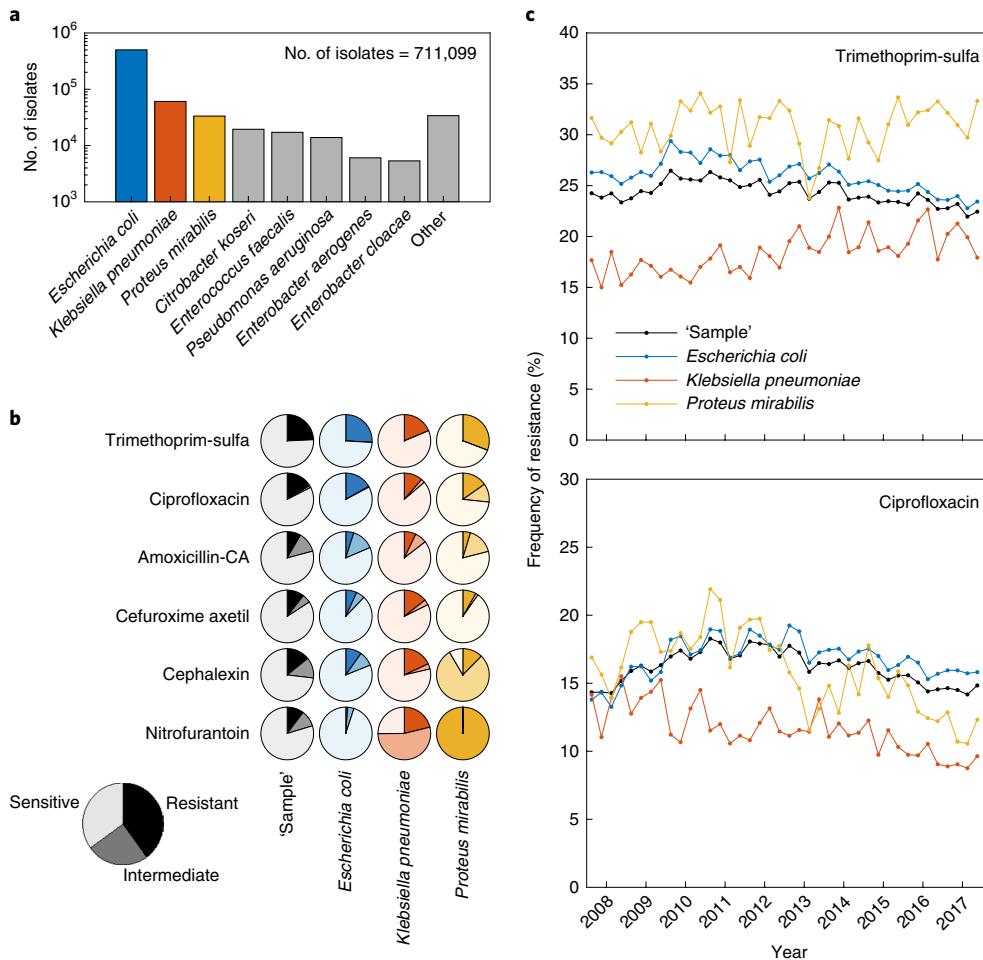
**Strong antibiotic-specific correlations of resistance with demographic factors.** Consistent with previous studies, UTIs were much

**Table 1 | List of antibiotic resistances analyzed in the study**

Antibiotics	Class
Trimethoprim-sulfa	Dihydrofolate reductase (DHFR) inhibitor
Ciprofloxacin	Fluoroquinolone
Amoxicillin-clavulanic acid (CA)	Penicillin-β-lactamase inhibitor
Cefuroxime axetil	Cephalosporin
Cephalexin	Cephalosporin
Nitrofurantoin	Nitrofuran

more common in females than males (~88% females)<sup>11,26</sup> and had qualitatively different age distributions (Fig. 2a)<sup>11,18,26,42,43</sup>. For each antibiotic, we performed multivariate logistic regression for the odds of resistance  $\eta = P_{\text{Resistance}}/(P_{\text{Sensitive}} + P_{\text{Intermediate}})$  as a function of age, gender, retirement home residence, pregnancy, date of sampling (time since July 2007) and season of sampling (Methods: Logistic regression 'Demographics' model; intermediate levels of resistance were classified as sensitive since they do not exclude prescription of an antibiotic, especially given the higher efficacy of antibiotics in urine infections<sup>44</sup>). We also calculated, for each of the six antibiotics, the frequencies of resistance of the urine samples across age, separated by gender, pregnancy and retirement home residence (Fig. 2c and Extended Data Fig. 3a).

Age, gender, pregnancy and residence in a retirement home had strong, yet differential, association with resistances to the six antibiotics. For all six antibiotics, the risk of resistance strongly increased with age and with retirement home residence and decreased for females and pregnancy (Fig. 2b,c; see Supplementary Table 2 for regression coefficients and 95% confidence intervals (CIs)). The odds ratio (OR) for age (the ratio between the adjusted odds of resistance in the oldest and youngest age groups; Methods) differed widely among the six measured antibiotics, ranging from 2 in trimethoprim-sulfa and amoxicillin-CA to >8 in ciprofloxacin (Fig. 2b and Supplementary Table 2). For some antibiotics, the risk of an infection being resistant were non-monotonic with age, having an additional peak of higher risk at infancy or childhood (for example, nitrofurantoin; Fig. 2c). For all antibiotics, females had lower odds of resistance, yet the ORs varied substantially among the different antibiotics (from OR = 0.95, 95% CI: 0.93–0.97 for trimethoprim-sulfa to OR = 0.38, 95% CI: 0.38–0.39 for cefuroxime axetil). These lower odds of resistance for females were often lowered even further with pregnancy (as much as OR = 0.48, 95% CI: 0.45–0.50 for ciprofloxacin; Supplementary Table 2). We also identified an interaction between gender and age leading to heterogeneous patterns for males and females (for example, trimethoprim-sulfa, nitrofurantoin) and even to opposing interactions of gender with specific age groups (for example, ciprofloxacin; Fig. 2c). While, across all antibiotics, resistance was higher for residents of retirement homes, the correlation with age within this group was reversed: the frequencies of resistance for retirement home residence did not increase, and even slightly decreased, with age (Fig. 2c and Extended Data Fig. 3a; possibly representing differential survivorship). The date of sample had some association with resistance to specific antibiotics, most notably cefuroxime axetil, while season had a relatively weak correlation with resistance for any of the drugs (Fig. 2b). Comparing the frequencies of resistance across the different antibiotics, we found that relative resistance rates changed between age groups (Extended Data Fig. 3b). We concluded that among the different demographic factors associated with risk of resistance, age, gender and residence in retirement homes were the strongest, with resistances to different antibiotics differentially correlated with these factors and the interactions among them.

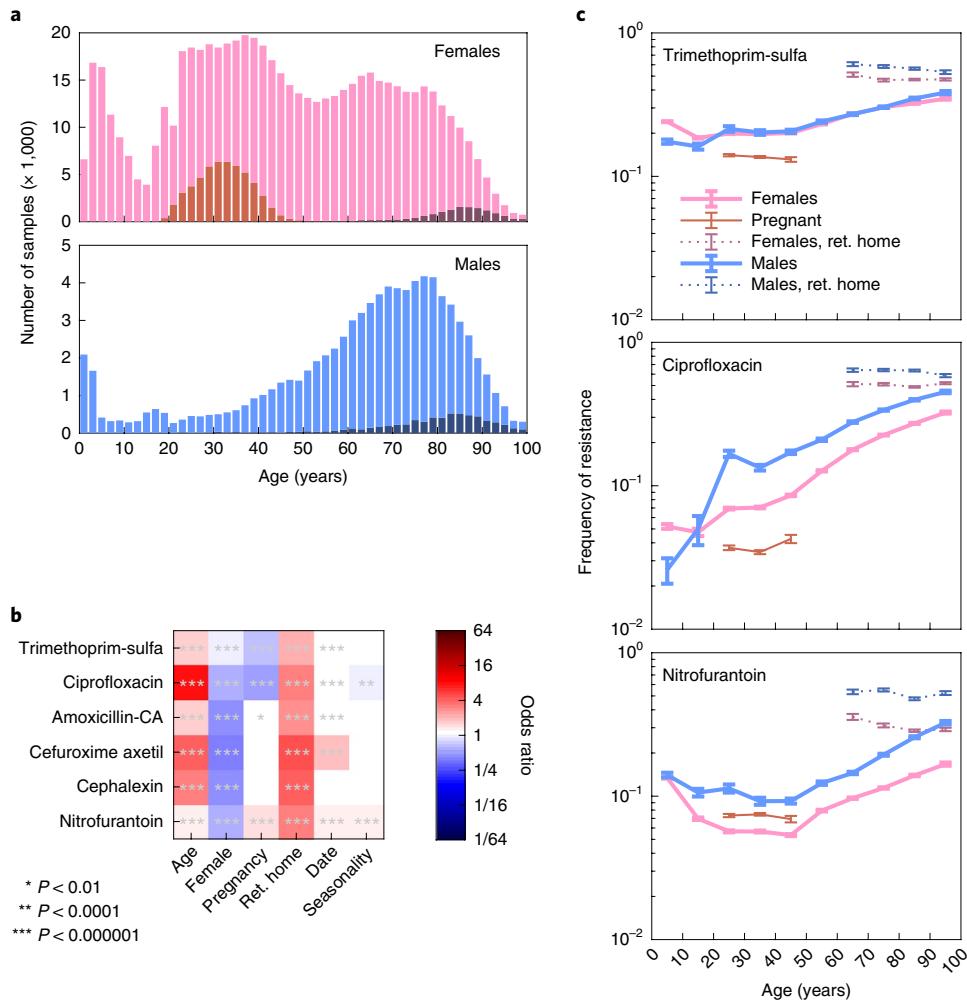


**Fig. 1 | Frequency of bacterial species and antibiotic resistance in urinary tract infections.** **a**, Species abundance across the entire UTI data set (July 2007–June 2017, 711,099 samples). **b**, The frequency of resistance and intermediate resistance to the six focal antibiotic drugs for the three most common bacterial species and for the urine sample as a whole ('sample', defined as the highest resistance measured for each isolate in the sample). Dark to light shades represent resistant, intermediate and sensitive, respectively. **c**, Frequencies of resistance for each of the three common species (colored lines) and the sample resistance (black lines) over the 10-year sampling time, for two representative antibiotics: trimethoprim-sulfa (top) and ciprofloxacin (bottom; see Extended Data Fig. 2 for all antibiotics). Data points represent quarterly averages.

**Long-term correlations of resistance among same-patient urine samples.** Moving from demographics to clinical history, we analyzed correlations of resistance across same-patient infections, revealing 'memory-like' long-term auto-correlations and a timeless patient-specific tendency for resistance. Analyzing all same-patient pairs of samples, we calculated for each antibiotic the risk ratio for resistance of the second sample given the resistance of the first sample ( $\zeta_{\text{pairs}} = [N_{R \rightarrow R}/(N_{R \rightarrow R} + N_{R \rightarrow S})]/[N_{S \rightarrow R}/(N_{S \rightarrow R} + N_{S \rightarrow S})]$ ), where  $N$  is the number of same-patient sample pairs with the specified resistance phenotypes; for example,  $N_{R \rightarrow S}$  is the number of sample pairs in which the first sample is resistant to the antibiotic and the second sensitive; Methods). Calculating  $\zeta_{\text{pairs}}$  as a function of the time difference  $t = t_1 - t_2$  between the two samples in each pair, we find that, for all antibiotics, these risk ratios are highest for short time differences and decay as the time difference increases (Fig. 3 and Supplementary Fig. 1). Sample pairs less than a week apart showed substantially higher risk ratios, which we interpreted as repeated measurements of the same infection (Supplementary Fig. 1). Considering only correlations between sample pairs more than a week apart, we found that the risk ratios decay and finally converge at long time differences, to an asymptotic constant larger than 1 (the risk ratios are well fitted by the sum of an exponent and a

constant,  $\zeta_{\text{pairs}} \approx C_m e^{t/\tau_m} + C_\infty$ ; Fig. 3a,b and Supplementary Fig. 1). The memory-like decay time  $\tau_m$  of correlations among samples was longer than 6 months for most antibiotics and even exceeded 1 year for ciprofloxacin resistance, which is consistent with and even longer than previously observed (Fig. 3c)<sup>34</sup>. The maximal risk ratios considering previous resistance reached about 8 for short time differences for some antibiotics and typically remained larger than 3 even for samples taken half a year apart (Fig. 3a,b and Supplementary Fig. 1). At much longer times, the risk ratio decayed, and  $\zeta_{\text{pairs}}$  converged to a constant, but interestingly it did not fully diminish, but rather converged to values larger than 1 (Fig. 3a,b,d, green), representing timeless patient-specific tendencies for resistance. These decaying memory-like and timeless correlations could stem from repeated same-strain infections or from correlations with other patient-specific factors. In either case, these strong memory-like and timeless correlations can potentiate predictions of resistance.

**Direct and indirect selection for resistance following past antibiotic purchase.** Next, we linked the infection data set with patient-resolved antibiotic purchase data. For each patient with recorded UTI samples, we retrieved all records of antibiotic purchase made during the 20-year period from 1 Jan 1998 to 30 Jun 2017. For analysis,



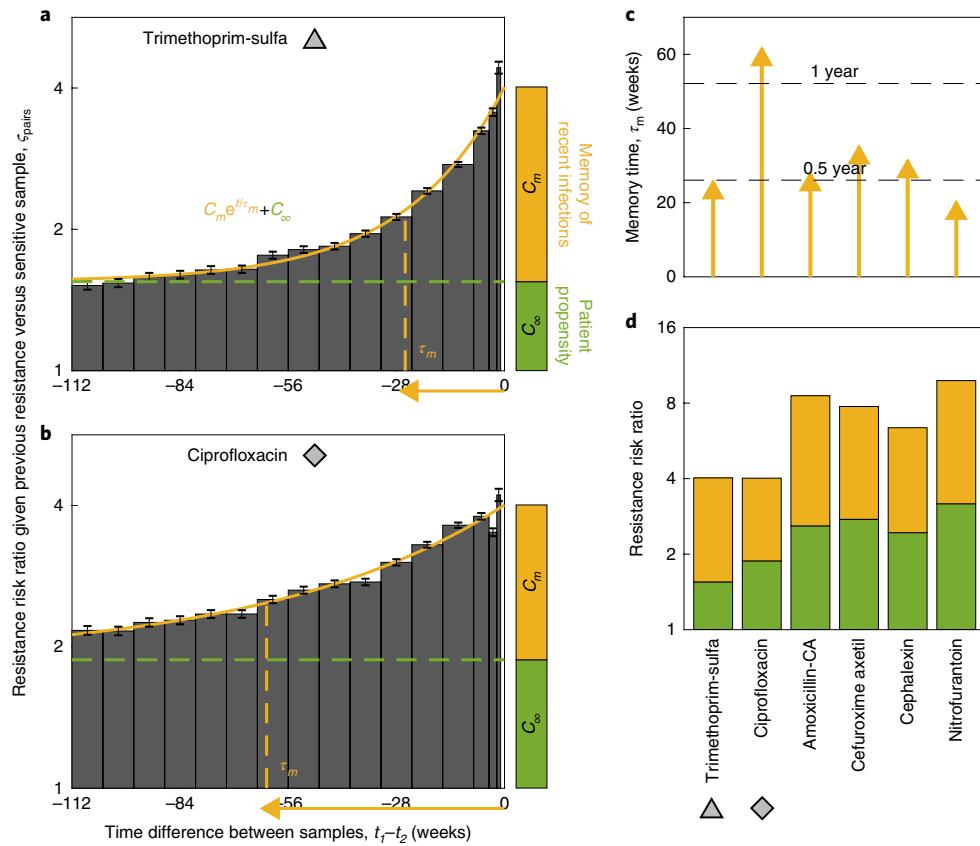
**Fig. 2 | Antibiotic-specific associations of resistance with demographic factors.** **a**, Distribution of urine cultures across major demographic factors: age, gender (top, females; bottom, males), pregnancy (red bars, top) and retirement (ret.) home residence (dark bars, top and bottom). **b**, Adjusted odds ratios of resistance for each demographic variable (see Methods, Logistic regression ‘Demographics’ model, and see Supplementary Table 2 for all adjusted and unadjusted regression coefficients). Asterisks indicate statistical significance and non-significant odds ratios ( $P > 0.01$ ) are shown as blank. **c**, Frequency of resistance as a function of age showing qualitatively distinct patterns for three representative antibiotics. UTI samples are separated into five non-overlapping categories: men not residing in retirement homes (blue), men residing in retirement homes (dotted blue), women not pregnant and not residing in retirement homes (magenta), women in retirement homes (magenta dotted) and pregnant women (red). See Extended Data Fig. 3 for all antibiotics.

we used the 11 most purchased drugs (Supplementary Table 1). Antibiotics identical or highly similar to the ones used for resistance measurement were assigned as cognate antibiotics of these resistance measurements (Methods; Supplementary Table 1). For each UTI sample, we counted the number of purchases made by the same patient of each of the 11 drugs at distinct time intervals before the sample (Methods). Then, we applied multivariate logistic regression to correlate resistance to each of the six antibiotics with these drug purchase counts (Methods: Logistic regression ‘Purchase history’; Fig. 4a and Extended Data Fig. 4a).

We identified strong long-term patient-level associations of resistance with past purchases of both cognate and non-cognate antibiotics. These purchase-resistance associations peaked at time differences of 1 to 2 weeks between purchase and sample, and often lasted for months and even longer than 1 year (Fig. 4a and Extended Data Fig. 4a). For example, the associations between the purchase of ciprofloxacin and its cognate resistance had an OR of 1.5 after half a year and remained as large as 1.2 even 2 years past the purchase (Fig. 4a). Some weak negative associations were also identified (for example, ciprofloxacin resistance was negatively correlated with the

past use of amoxicillin and cefalexin; Fig. 4a). Yet, the magnitude of these negative correlations decreased after adjusting for demographics, suggesting that they stemmed indirectly from correlations of purchases and resistance with demographics (Methods: Logistic regression, ‘Purchase history adjusted for demographics’; Extended Data Fig. 4c). Notably, drug purchases were associated not only with their expected cognate resistances. Indeed, use of some first-line antibiotics, such as ciprofloxacin and ofloxacin, increased the risk of a future resistance to a wide range of mechanistically diverse antibiotics. These abundant long-term positive associations between resistances and past purchase of non-cognate drugs did not stem from correlations of purchases and resistance with patient demographics; they remained strong even when adjusting for demographics (Extended Data Fig. 4c). Together, these results support strong and long-lasting patient-level associations of antibiotic resistance with the past use of both cognate and non-cognate antibiotics.

Exposing direct drug-to-resistance associations by disentangling correlations among resistances, we found that drug usage specifically selects for its cognate resistance at the single patient level. Across the sample data set, resistances to different antibiotics within class



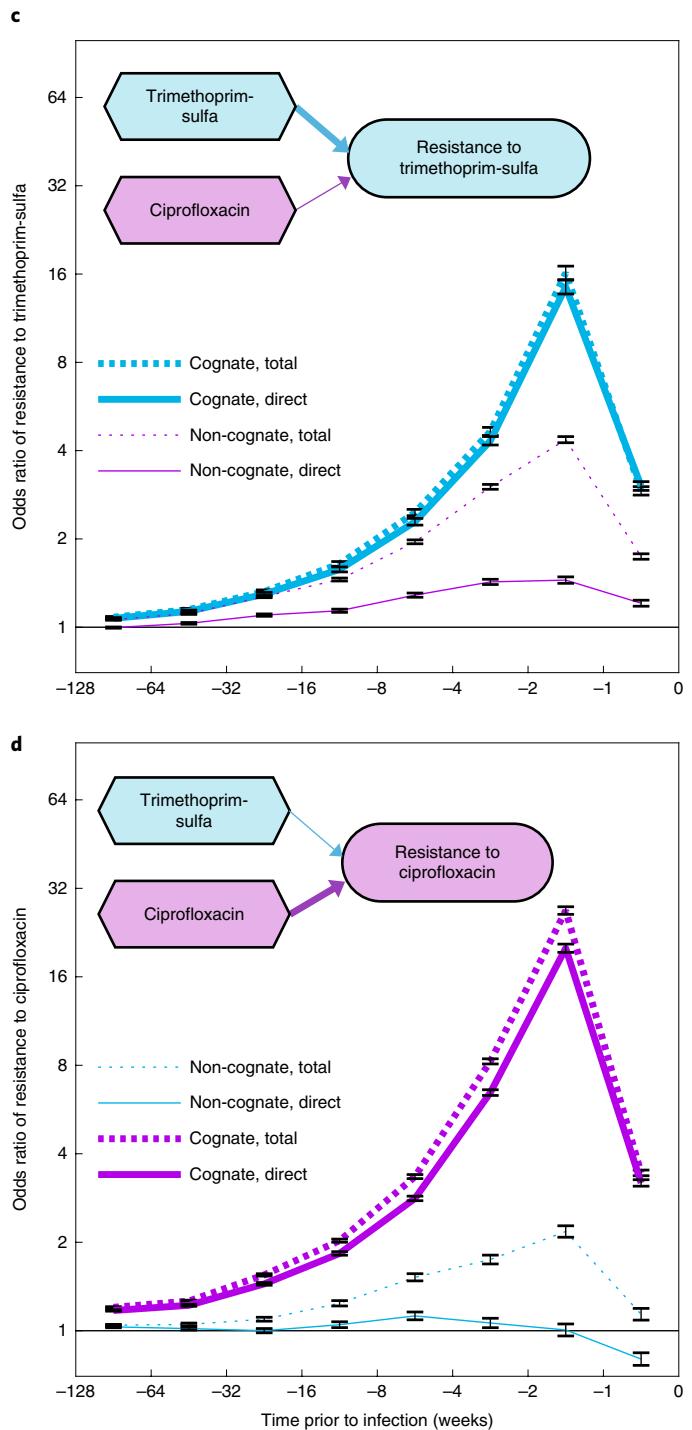
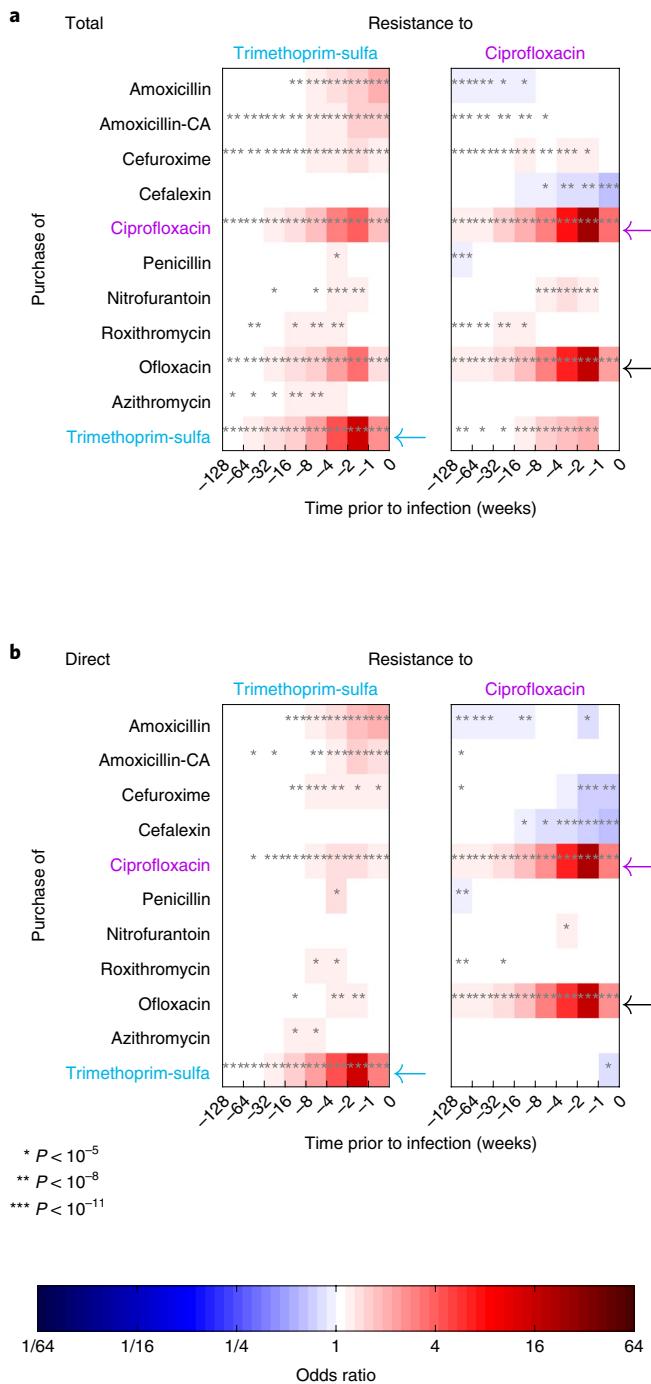
**Fig. 3 | Long-term ‘memory’ of resistance across same-patient samples.** **a,b**, Risk ratio of the resistance of a urine sample given a record of a resistant versus sensitive earlier sample from the same patient, as a function of the time difference between the two samples, for trimethoprim-sulfa (**a**) and ciprofloxacin (**b**, see Methods and Supplementary Fig. 1 for all antibiotics). Risk ratios are well fitted with  $\zeta_{\text{pairs}} \approx C_m e^{t/\tau_m} + C_\infty$  representing a time-decaying correlation (‘memory’, yellow) and a time-independent correlation (‘patient propensity’, green) among sample pairs. The magnitudes of these terms are shown as stacked bars on the right and the memory time ( $\tau_m$ ) is indicated across the time axis (yellow arrow). Gray triangles and diamonds represent trimethoprim-sulfa and ciprofloxacin, respectively, linking between the different panels. **c**, Time scales of the memory of resistance  $\tau_m$  for the six different antibiotics (correlated with the yellow arrows in **a** and **b**). **d**, The magnitude of long-term and timeless memory for the different antibiotics (yellow and green bars, respectively).

and even resistances to antibiotics of different classes were highly correlated (cross-resistance; Extended Data Fig. 5). These inherent correlations among resistances suggest that observed associations between resistance to a given drug A and past purchase of a different non-cognate drug B may arise indirectly through selection for resistance B and association between resistance to B and resistance to A. Mathematically discerning these direct and indirect effects is only possible when multiple resistances are considered<sup>20,45</sup>. As our data set contained measurements of multiple resistances for each sample, we were able to disentangle direct from indirect associations by adjusting the logistic regression for other measured resistances (Methods: Logistic regression ‘Purchase history adjusted for cross-resistance’). In this cross-resistance-adjusted analysis of purchase-resistance associations, the non-cognate associations between drug purchases and resistance substantially diminished and even disappeared, while the associations between cognate drug-to-resistance pairs persisted (Fig. 4b and Extended Data Fig. 4b). For example, considering the associations between purchases of trimethoprim-sulfa and ciprofloxacin to their cognate resistances, we observed that the unadjusted and cross-resistance-adjusted associations were of similar magnitude for cognate drugs (Fig. 4c,d, thick solid versus thick dashed lines), while the total association of drugs with their non-cognate resistance decreased considerably once the indirect effect was removed (Fig. 4c,d, thin solid versus thin dashed lines). Our analysis therefore identifies both direct and indirect selection

for resistance at the single patient level lasting months and even 1 year following drug use.

**Predicting antibiotic resistance at the single-patient single-infection level.** As resistance is strongly associated with demographics, sample history and purchase history, we wanted to determine the predictive power of these factors individually and when combined together and identify potential interactions among them. Models of logistic regression and gradient-boosting decision trees (GBDTs) were trained and tested on temporally separate periods: a training period of 9 years from 1 July 2007 to 30 June 2016 and a testing period of the following year, from 1 July 2016 to 30 June 2017 (for cephalaxin, the training period was modified to avoid a time period during which resistance to this drug was not routinely measured, Extended Data Fig. 1). This temporal separation between training and testing data emulates forecasting resistance, as would be the case in real-life implementation of such a method. The area under the curve (AUC) of receiver operating characteristic was used as a standard measure for predictive power<sup>46</sup>.

Logistic regression and GBDT models provided personalized drug-specific prediction of resistance. Individually considering demographics, sample history and purchase history, we find that each of these sets of features had significant predictive power, with their relative prominence varying across the different antibiotics (Extended Data Fig. 6). Combining all these feature sets in



**Fig. 4 | Direct association of past purchase with its cognate resistance leads, through association among resistances, to indirect association of purchases with non-cognate resistances.** **a**, Multivariate logistic regression models for the association of resistance to trimethoprim-sulfa (left) and ciprofloxacin (right) with past purchases of the indicated drugs at the indicated time intervals before infection ('Total', see Extended Data Fig. 4a for all antibiotics; Logistic regression—purchase history in Methods). Values represent the odds ratios for a single purchase of a specific drug at a specific time interval (color map, stars for statistical significance as indicated, non-significant values, with Bonferroni corrected  $P > 0.05$ , are blanked). A long-term association is observed between resistance and past purchase of its matching (cognate, arrows) as well as with non-cognate antibiotics. **b**, Logistic regression model as in **a** adjusted for cross-resistance. This adjusted model diminishes or even completely removes non-cognate drug-to-resistance associations while fully preserving the cognate associations ('Direct', see Extended Data Fig. 4b for all antibiotics; arrows; cyan, trimethoprim-sulfa; magenta, ciprofloxacin). **c,d**, Association of resistance to trimethoprim-sulfa (**c**) and ciprofloxacin (**d**) with purchases of these two drug (cyan and magenta, respectively). Note differences between total (dashed lines) and direct (solid lines) effects for cognate (thick lines) versus non-cognate (thin lines) drugs.

a complete logistic regression model (Methods: Logistic regression 'Complete'), greatly increased the predictability of resistance (AUC ranged from 0.7 for amoxicillin-CA to 0.83 for ciprofloxacin;

Extended Data Fig. 6). Predictability of resistance was slightly increased by the GBDT models (Methods). For each given antibiotic  $k$ , considering the model-assigned resistance probabilities  $P_k^m$

of each sample  $m$ , we can define threshold values  $P_k^{\text{threshold}}$  that allow substantial reduction in the risk of resistance, while allowing treatment of the vast majority of the infections (Fig. 5a). Setting this threshold to allow treatment of 75% of samples by each of the six drugs, the vast majority of infections can be treated with at least one of the drugs (92%, Extended Data Fig. 7). Finally, we found that these model-assigned probabilities of resistance can markedly differentiate samples resistant to one drug and sensitive to another (Fig. 5b, OR of 3.9 for nitrofurantoin versus cefuroxime axetil,  $P < 10^{-100}$ , Fisher's exact test; see Supplementary Fig. 3 for all other drug pairs). In total, these results demonstrate that machine-learning models can provide high and specific predictability of antibiotic resistance at the single-patient and single-infection levels, motivating the development of algorithmic drug recommendations and comparison of their performance with current standard of care.

**Algorithmic drug recommendations substantially reduce mismatched treatments.** Analyzing prescriptions given by physicians as part of current standard of care, we found that these prescriptions significantly, yet not strongly, reduced the rate of mismatched treatments, compared with null random expectations. We identified all cases of 'same-day empirical treatments' (SDETs), where a patient purchased an antibiotic on the same day they had a UTI sample sent for culture (11,952 cases within the 1-year test period; since culture tests take 2–4 d, these prescriptions were necessarily given empirically). Retrospectively contrasting these empirically prescribed drugs with the measured resistance of their corresponding samples, we found an overall 8.5% (95% CI: 8.03–9.05) rate of mismatched treatments (the sample was resistant to the prescribed antibiotic). This rate was significantly, yet not strongly, lower than expected by chance in two different null models. First, randomly choosing for each of these SDET cases, one of the six drugs with equal probabilities, we found an expected null mismatched treatment rate of 10.2% (95% CI: 9.88–10.52), which is 20% higher than observed in physicians' prescriptions ( $P < 10^{-10}$ , Bootstrapping, Methods; Fig. 5c). Second, randomly permuting among the SDET cases the same pool of drugs prescribed by the physicians, we found an expected null mismatched rate of 9.4% (95% CI: 9.00–9.71), namely 10% higher than observed ( $P = 2.3 \times 10^{-5}$ , Bootstrapping, Methods; Fig. 5c). Together, these results indicate statistically significant, but mild, patient-specific optimization of treatment in standard clinical practice.

Developing algorithmic drug recommendations based on the machine-learning predictions of resistance, we found that they can greatly improve on these standard-of-care rates of mismatched empirical treatments. To computationally recommend drugs based on the machine-learning assigned probabilities of resistance  $P_k^m$ , we considered two algorithms, unconstrained and constrained (cost-adjusted; Extended Data Fig. 8). In the unconstrained model, we simply chose for each of the SDET cases the antibiotic for which the model predicted the risk of resistance was lowest (minimal  $P_k^m$ , 'Unconstrained algorithm for drug choice', Methods). Comparing these recommendations to the measured antibiotic susceptibility of the sample, we found a mismatched rate as low as 5.1% (95% CI: 4.69–5.48) namely 42% lower than observed in the physician-prescribed treatment of the same cases ( $P < 10^{-10}$ , Bootstrapping, Methods; Fig. 5c). The chance of mismatched treatment was lower than expected not only in total, but across each of the prescribed drugs (Fig. 5d, top). Importantly though, the distribution of drugs recommended by this unconstrained algorithm was very different to the distribution of drugs prescribed by physicians (Fig. 5d, bottom). In particular, the algorithm almost entirely refrained from prescribing trimethoprim and cefalexin, for which population-level rates of resistance were high. Optimal unconstrained algorithmic recommendations can thus dramatically reduce the chance of mismatched treatments, yet do so by drastically changing the overall distribution of prescribed drugs.

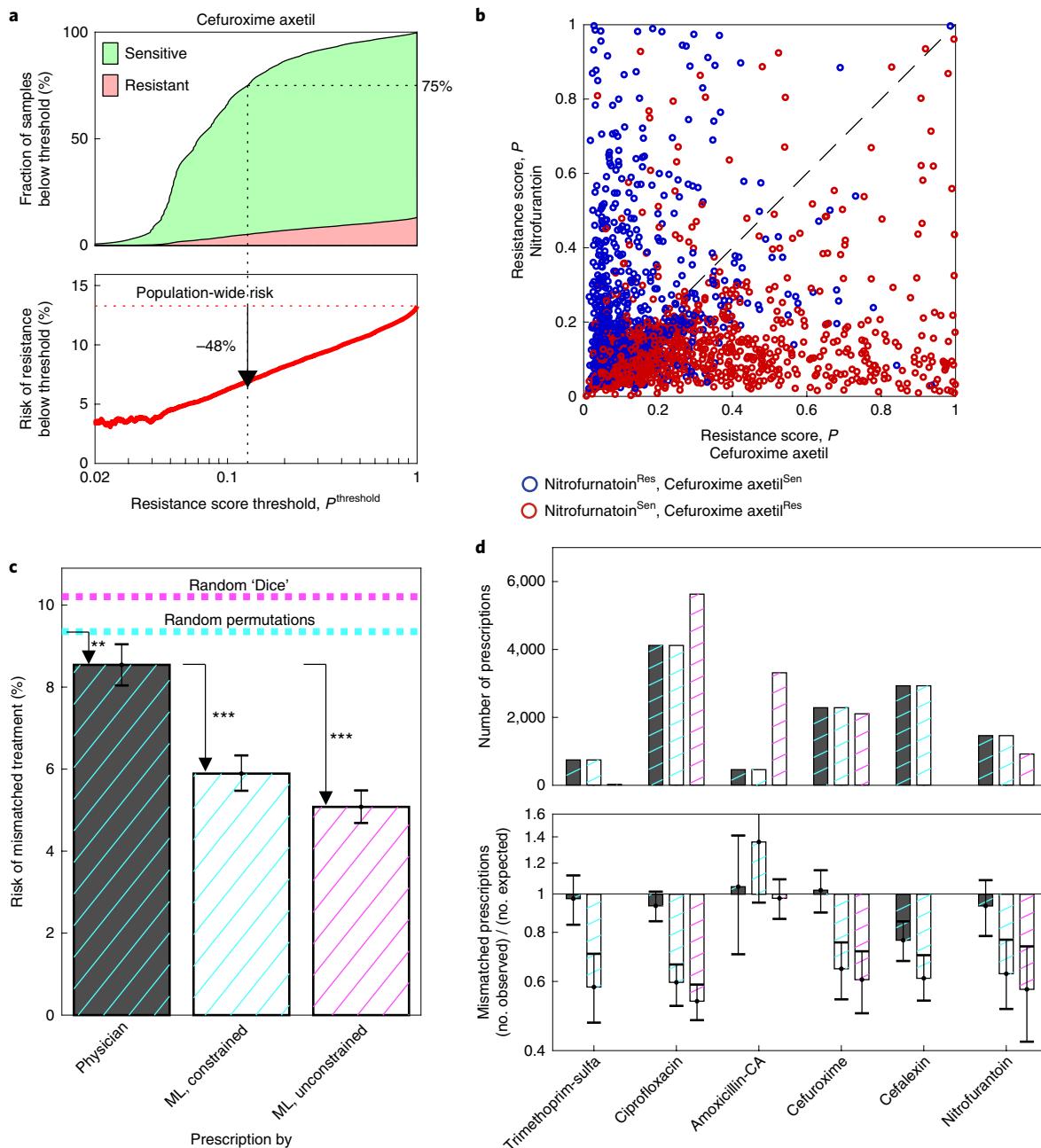
A model constrained to prescribe each drug at the same frequency used by physicians can still greatly reduce the rate of mismatched treatments. The overall rate of prescription of each drug could reflect considerations other than minimizing mismatched treatment (for example, ease of use, side effects and tendency to avoid drugs for which population-level resistance rates are low). To address these considerations, here referred to as costs, we developed a constrained, cost-adjusted, algorithm ('Constrained (cost-adjusted) algorithm for drug choice', Methods). To recommend drugs that best minimize the population rate of mismatched treatments while maintaining a given population-level frequency of use of each drug, the algorithm assigns an effective cost for each drug and adjusts their values to match the required distribution of drug use (Methods). Applying this model to the SDET cases while adjusting the drug-specific costs, such that the overall distribution of recommended drugs precisely matches the distribution of the drugs prescribed by physicians, this model gave a mismatched treatment rate of 5.9% (95% CI: 5.47–6.33), slightly above the unconstrained model but still 30% lower than the physicians' rate ( $P < 10^{-10}$ , Bootstrapping, Methods). The improvements in mismatch rate were general across the population and robust to the clinical definition of resistance (Extended Data Fig. 9). These results show that algorithmically suggested drug prescriptions can substantially reduce the risk of mismatched treatments even when allowed to barely permute the same pool of drugs among patients.

## Discussion

Analyzing a large longitudinal medical data set, we demonstrate a high predictability of antibiotic resistance in UTIs, which can guide culture-free recommendation of treatment to lower the chance of mismatched empirical treatment. The best predictive power of resistance comes from combining patient-specific data of demographics, antibiotic resistance profiles of past UTIs and purchase history of antibiotic drugs. Considering demographics, we found that age, gender, pregnancy and residence in a retirement home were strongly associated with resistance, showing complex and non-monotonic patterns specific to each of the different antibiotics. Utilizing repeated same-patient cultures in our database, we identified and characterized a personal component of memory-like correlations of resistance, lasting for many months and even over 1 year. These long-term correlations can represent recurrent infections with the same strain or correlations with other patient-specific factors. Either way, we showed that they further contribute to predictability of resistance.

Long-term associations were also observed between resistance and past drug purchases. Resistance to a given drug had long-lasting associations not only with past usage of this same drug, but also with other, even mechanistically unrelated, drugs. Yet, adjusting for correlations among resistances exposed direct selection, where drug use led specifically to its own cognate resistance at the single-patient level. These results are consistent with drug use directly selecting, at the single-patient level, for strains resistant to it and thereby selecting indirectly, likely through frequent co-occurrence, to resistance to other antibiotics.

Combining these demographic, sample history and drug history data can guide algorithmic recommendations for empirical treatment that substantially improve on the current standard of care. Comparing empirical prescriptions given by physicians to random prescriptions, we found that physicians personalize drug prescriptions in ways that significantly reduce the chance of mismatched treatment. However, machine-learning models could still substantially improve on these already reduced rates. Indeed, the rates of mismatched treatment would have been reduced by over 40% were the drugs with lowest machine-learning predicted chance of resistance chosen. These machine-learning recommendations are inherently biased towards recommending drugs with overall low levels of



**Fig. 5 | Algorithmically suggesting antibiotic prescription for empirical treatments can greatly improve upon the current standard of care.** **a**, For each of the six antibiotics, we calculated the fraction (top) of resistant (red) and sensitive (green) samples, as well as the risk of resistance (bottom), for all samples within the 1-year test period whose complete-model machine-learning assigned probabilities of resistance  $P_k^m$  were below a set threshold  $P_{\text{threshold}}$  ( $x$  axis; see Supplementary Fig. 2 for all antibiotics and more formal definitions). At  $P_{\text{threshold}}=1$  the risk of sample resistance equals the population-wide risk of resistance (dotted red line). Setting  $P_{\text{threshold}}=0.12$  would permit treatment of 75% of these infections with much reduced risk of resistance compared with the population-wide risk (48% reduction, down-pointing arrow). **b**, Differentiation between samples resistant to cefuroxime axetil and sensitive to nitrofurantoin (red) and vice versa (blue) by their model-assigned resistance probabilities (OR of 3.9 for red points below the diagonal and blue points above it;  $P < 10^{-100}$ , Fisher's exact test; see Supplementary Fig. 3 for all pairs of antibiotics). **c**, Physician's frequency of mismatched prescriptions across all SDET cases (dark bar) was slightly better than the null expectation for randomly prescribing drugs with equal probabilities (random 'dice', magenta dashed,  $P < 10^{-10}$ ) or for randomly permuting the physicians' prescriptions (random permutations, cyan dashed,  $P = 2.5 \times 10^{-5}$ ). These mismatch treatment rates were substantially reduced by the machine-learning (ML)-based recommendations (light bars), either unconstrained (magenta hatched,  $P < 10^{-10}$ ) or constrained to recommend drugs at the exact same frequencies prescribed by the physicians (cyan hatched,  $P < 10^{-10}$ ). **d**, Top, distribution of the drugs prescribed by the physicians (dark bar), by the constrained algorithm (cyan-hashed light bar, constrained to be equal to the physician's) and by the unconstrained algorithm (magenta-hashed light bar). Bottom, for each of these prescription models, the frequency of mismatched treatment for each of the drugs is indicated, normalized by the expected mismatch frequency for random drug prescription (the average rate of resistance to the drug across the SDET population).

resistance, for example ciprofloxacin, which is often intentionally avoided in standard clinical practice precisely to hinder the spread of resistance. We therefore also developed a model that assigns a

cost for each drug, thereby constraining the rate of recommendation of each drug to the rate at which it was prescribed by physicians. Importantly, even when constrained to merely permute among the

patients the exact same pool of drugs prescribed by physicians, the model could still reduce the rate of unmatched treatment by over 30% compared with standard care.

Some aspects of the data may complicate the interpretation of our results. As purchase of a drug does not fully guarantee its concurrent use, later usage of a purchased drug may bias our results towards a higher OR for purchases made long before infection. Conversely, we can not exclude that some patients have used antibiotics they did not purchase through MHS, which will bias our results towards lower ORs for drug purchases. Additionally, past antibiotic purchase and treatment might be associated with different clinical conditions, not considered in this study, such as comorbidities, hospitalizations and catheter use. While these factors are less likely to directly affect resistance rates, they are likely associated with risk of infections. Also, although culture data are routine for suspected UTIs, sending urine for a culture test is not obligatory. As a result, we assume some UTIs would be empirically treated without any culture record, and there is likely higher propensity towards culture testing of infections suspected of being resistant. This would generate bias towards the measurement of more resistant samples, resulting in an overestimation of the total frequency of resistance, especially for first-line treatment and potentially in overestimation of the general rate of mismatched treatment. Another bias due to elective culture testing would be for cultures taken following treatment failure. Such bias can again generate bias towards measurements of more resistant samples and it can further contribute to the strong short-term association of drug purchases with resistance, especially for first-line antibiotics. Lastly, the extent of this bias towards culture testing specifically following treatment failure could itself depend on demographics, which can bias correlations of demographics with resistance. While we can not exclude these biases, our analysis demonstrates that, with all of these potential biases, resistance of urine infections can be well predicted based on the specific demographics and clinical history of the patient, and that algorithmic drug recommendations can substantially reduce the chance of prescribing an antibiotic to which the infection is resistant.

The substantial reduction in the rate of mismatched treatment enabled by machine-learning recommendations based on the patient's record and clinical history lays the basis for a future paradigm where clinicians will routinely consult such algorithms for prescription of patient-tailored antibiotic treatment. We expect that algorithmic approaches similar to the one described here will be implemented, either centralized or locally, in healthcare systems where vast longitudinal electronic health records are available. While the key factors identified here can serve as the basis of such an approach, the specific model, the exact coefficients and relative weights of predictors, will have to be adjusted for each country or region. Indeed, these algorithms can also be dynamically and adaptively updated in real time as new data are acquired. We expect that inclusion of additional patient-specific factors, such as comorbidities and hospitalizations, as well as of real-time information on infections, resistance and drug usage in other patients in a range of geographical proximities<sup>39</sup>, can further increase resistance predictability. These models could also be used to adjust for patient-specific drug 'costs', thereby accounting for allergies and other patient-specific drug restrictions. In the longer term, these clinical-record and epidemiological data-based approaches could be integrated with genomics of the patient as well as of the pathogen<sup>47–53</sup>. Implemented in the clinic, machine-learning-guided personalized empirical prescription can reduce treatment failure as well as lower the overall use of antibiotics, thereby assisting in the global effort of impeding the antibiotic resistance epidemic.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and

associated accession codes are available at <https://doi.org/10.1038/s41591-019-0503-6>.

Received: 6 August 2018; Accepted: 30 May 2019;  
Published online: 4 July 2019

## References

- Ventola, C. L. The antibiotic resistance crisis: part 1: causes threats. *P T.* **40**, 277–283 (2015).
- Rossolini, G. M., Arena, F., Pecile, P. & Pollini, S. Update on the antibiotic resistance crisis. *Curr. Opin. Pharmacol.* **18**, 56–60 (2014).
- Goossens, H., Ferech, M., Vander Stichele, R. & Elseviers, M., ESAC Project Group. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* **365**, 579–587 (2005).
- Bronzwaer, S. L. A. M. et al. A European study on the relationship between antimicrobial use and antimicrobial resistance. *Emerg. Infect. Dis.* **8**, 278–282 (2002).
- Costelloe, C., Metcalfe, C., Lovering, A., Mant, D. & Hay, A. D. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *BMJ* **340**, c2096 (2010).
- Fridkin, S. K. et al. The effect of vancomycin and third-generation cephalosporins on prevalence of vancomycin-resistant enterococci in 126 U.S. adult intensive care units. *Ann. Intern. Med.* **135**, 175–183 (2001).
- Malhotra-Kumar, S., Lammens, C., Coenen, S., Van Herck, K. & Goossens, H. Effect of azithromycin and clarithromycin therapy on pharyngeal carriage of macrolide-resistant streptococci in healthy volunteers: a randomised, double-blind, placebo-controlled study. *Lancet* **369**, 482–490 (2007).
- Kang, C.-I. et al. Bloodstream infections caused by antibiotic-resistant Gram-negative bacilli: risk factors for mortality and impact of inappropriate initial antimicrobial therapy on outcome. *Antimicrob. Agents Chemother.* **49**, 760–766 (2005).
- Kumar, A. et al. Initiation of inappropriate antimicrobial therapy results in a fivefold reduction of survival in human septic shock. *Chest* **136**, 1237–1248 (2009).
- Huang, A. M. et al. Impact of rapid organism identification via matrix-assisted laser desorption/ionization time-of-flight combined with antimicrobial stewardship team intervention in adult patients with bacteremia and candidemia. *Clin. Infect. Dis.* **57**, 1237–1245 (2013).
- Stamm, W. E. & Norrby, S. R. Urinary tract infections: disease panorama and challenges. *J. Infect. Dis.* **183**, S1–4 (2001).
- Engel, J. D. & Schaeffer, A. J. Evaluation of and antimicrobial therapy for recurrent urinary tract infections in women. *Urol. Clin. North. Am.* **25**, 685–701 (2005).
- Geerlings, S. E. Clinical presentations and epidemiology of urinary tract infections. *Microbiol. Spectr.* **4** <https://doi.org/10.1128/microbiolspec.UTI-0002-2012> (2016).
- Shapiro, D. J., Hicks, L. A., Pavia, A. T. & Hersh, A. L. Antibiotic prescribing for adults in ambulatory care in the USA, 2007–09. *J. Antimicrob. Chemother.* **69**, 234–240 (2014).
- Low, M. et al. Infectious disease burden and antibiotic prescribing in primary care in Israel. *Ann. Clin. Microbiol. Antimicrob.* **17**, 26 (2018).
- Kahlmeter, G. An international survey of the antimicrobial susceptibility of pathogens from uncomplicated urinary tract infections: the ECO/textperiodcentered SENS Project. *J. Antimicrob. Chemother.* **51**, 69–76 (2003).
- Farrell, D. J., Morrissey, I., De Rubis, D., Robbins, M. & Felmingham, D. A UK multicentre study of the antimicrobial susceptibility of bacterial pathogens causing urinary tract infection. *J. Infect.* **46**, 94–100 (2003).
- Foxman, B. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. *Am. J. Med.* **113**, 5S–13S (2002).
- Flores-Mireles, A. L., Walker, J. N., Caparon, M. & Hultgren, S. J. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat. Rev. Microbiol.* **13**, 269–284 (2015).
- Pouwels, K. B. et al. Association between use of different antibiotics and trimethoprim resistance: going beyond the obvious crude association. *J. Antimicrob. Chemother.* **73**, 1700–1707 (2018).
- Ashkenazi, S., Even-Tov, S., Samra, Z. & Dinari, G. Uropathogens of various childhood populations and their antibiotic susceptibility. *Pediatr. Infect. Dis. J.* **10**, 742–746 (1991).
- Kahan, N. R. et al. Empiric treatment of uncomplicated urinary tract infection with fluoroquinolones in older women in Israel: another lost treatment option? *Ann. Pharmacother.* **40**, 2223–2227 (2006).
- Hooton, T. M., Besser, R., Foxman, B., Fritsche, T. R. & Nicolle, L. E. Acute uncomplicated cystitis in an era of increasing antibiotic resistance: a proposed approach to empirical therapy. *Clin. Infect. Dis.* **39**, 75–80 (2004).
- Arslan, H., Azap, O. K., Ergönül, O. & Timurkaynak, F. Urinary Tract Infection Study Group. Risk factors for ciprofloxacin resistance among Escherichia coli strains isolated from community-acquired urinary tract infections in Turkey. *J. Antimicrob. Chemother.* **56**, 914–918 (2005).

25. Ikram, R., Psutka, R., Carter, A. & Priest, P. An outbreak of multi-drug resistant *Escherichia coli* urinary tract infection in an elderly population: a case-control study of risk factors. *BMC Infect. Dis.* **15**, 224 (2015).
26. Foxman, B. & Brown, P. Epidemiology of urinary tract infections: transmission and risk factors, incidence, and costs. *Infect. Dis. Clin. North Am.* **17**, 227–241 (2003).
27. Tenney, J., Hudson, N., Alnifaify, H., Li, J. T. C. & Fung, K. H. Risk factors for acquiring multidrug-resistant organisms in urinary tract infections: a systematic literature review. *Saudi Pharm J.* **26**, 678–684 (2018).
28. Colgan, R., Johnson, J. R., Kuskowski, M. & Gupta, K. Risk factors for trimethoprim-sulfamethoxazole resistance in patients with acute uncomplicated cystitis. *Antimicrob. Agents Chemother.* **52**, 846–851 (2008).
29. Burman, W. J. et al. Conventional and molecular epidemiology of trimethoprim-sulfamethoxazole resistance among urinary *Escherichia coli* isolates. *Am. J. Med.* **115**, 358–364 (2003).
30. Kang, M.-S., Lee, B.-S., Lee, H.-J., Hwang, S.-W. & Han, Z.-A. Prevalence of and risk factors for multidrug-resistant bacteria in urine cultures of spinal cord injury patients. *Ann. Rehabil. Med.* **39**, 686–695 (2015).
31. Lee, G., Cho, Y.-H., Shim, B. S. & Lee, S. D. Risk factors for antimicrobial resistance among the *Escherichia coli* strains isolated from Korean patients with acute uncomplicated cystitis: a prospective and nationwide study. *J. Korean Med. Sci.* **25**, 1205–1209 (2010).
32. Johnson, L. et al. Emergence of fluoroquinolone resistance in outpatient urinary *Escherichia coli* isolates. *Am. J. Med.* **121**, 876–884 (2008).
33. Paul, M. et al. Improving empirical antibiotic treatment using TREAT, a computerized decision support system: cluster randomized trial. *J. Antimicrob. Chemother.* **58**, 1238–1245 (2006).
34. MacFadden, D. R., Ridgway, J. P., Robicsek, A., Ellingsen, M. & Daneman, N. Predictive utility of prior positive urine cultures. *Clin. Infect. Dis.* **59**, 1265–1271 (2014).
35. Olesen, S. W., Barnett, M. L., MacFadden, D. R., Lipsitch, M. & Grad, Y. H. Trends in outpatient antibiotic use and prescribing practice among US older adults, 2011–15: observational study. *BMJ* **362**, k3155 (2018).
36. Ena, J., Amador, C., Martinez, C. & Ortiz de la Tabla, V. Risk factors for acquisition of urinary tract infections caused by ciprofloxacin resistant *Escherichia coli*. *J. Urol.* **153**, 117–120 (1995).
37. Brown, P. D., Freeman, A. & Foxman, B. Prevalence and predictors of trimethoprim-sulfamethoxazole resistance among uropathogenic *Escherichia coli* isolates in Michigan. *Clin. Infect. Dis.* **34**, 1061–1066 (2002).
38. Metlay, J. P., Strom, B. L. & Asch, D. A. Prior antimicrobial drug exposure: a risk factor for trimethoprim-sulfamethoxazole-resistant urinary tract infections. *J. Antimicrob. Chemother.* **51**, 963–970 (2003).
39. Low, M. et al. Association between urinary community-acquired fluoroquinolone-resistant *Escherichia coli* and neighbourhood antibiotic consumption: a population-based case-control study. *Lancet Infect. Dis.* **19**, 419–428 (2019).
40. Wang, A., Daneman, N., Tan, C., Brownstein, J. S. & MacFadden, D. R. Evaluating the relationship between hospital antibiotic use and antibiotic resistance in common nosocomial pathogens. *Infect. Control Hosp. Epidemiol.* **38**, 1457–1463 (2017).
41. Gupta, K. et al. International clinical practice guidelines for the treatment of acute uncomplicated cystitis and pyelonephritis in women: a 2010 update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. *Clin. Infect. Dis.* **52**, e103–e120 (2011).
42. Lipsky, B. A. Urinary tract infections in men. Epidemiology, pathophysiology, diagnosis, and treatment. *Ann. Intern. Med.* **110**, 138–150 (1989).
43. Ginsburg, C. M. & McCracken, G. H. Jr. Urinary tract infections in young infants. *Pediatrics* **69**, 409–412 (1982).
44. Edlin, R. S., Shapiro, D. J., Hersh, A. L. & Copp, H. L. Antibiotic resistance patterns of outpatient pediatric urinary tract infections. *J. Urol.* **190**, 222–227 (2013).
45. Kahlmeter, G. & Menday, P. Cross-resistance and associated resistance in 2478 *Escherichia coli* isolates from the Pan-European ECO/textperiodcentered SENS Project surveying the antimicrobial susceptibility of pathogens from uncomplicated urinary tract infections. *J. Antimicrob. Chemother.* **52**, 128–131 (2003).
46. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
47. Lieberman, T. D. et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* **43**, 1275–1280 (2011).
48. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* **13**, 601–612 (2012).
49. Bradley, P. et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun.* **6**, 10063 (2015).
50. Khoury, M. J. & Ioannidis, J. P. A. Medicine. Big data meets public health. *Science* **346**, 1054–1055 (2014).
51. Beam, A. L. & Kohane, I. S. Big data and machine learning in health care. *JAMA* **319**, 1317–1318 (2018).
52. Grad, Y. H. & Lipsitch, M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol.* **15**, 538 (2014).
53. Sandora, T. J., Gerner-Smidt, P. & McAdam, A. J. What's your subtype? The epidemiologic utility of bacterial whole-genome sequencing. *Clin. Chem.* **60**, 586–588 (2014).

## Acknowledgments

We thank M. Datta, A. McAdam, G. Priebe and P. Ramesh for thorough reading of the manuscript and important comments. This work was supported in part by US National Institutes of Health grant no. R01 GM081617 (to R.K.) and European Research Council FP7 ERC grant no. 281891 (to R.K.) as well as The Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine (to R.K.).

## Author contributions

V.S. and R. Kishony perceived the study. I.Y., O.S., G.K., V.S. and R. Kishony designed the study. R. Katz, M.P., O.T. and V.S. retrieved and interpreted electronic health records. I.Y., O.S., G.N. and R. Kishony analyzed the data. I.Y., O.S., G.C., V.S. and R. Kishony interpreted the results. I.Y. and R. Kishony wrote the manuscript with comments from all authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41591-019-0503-6>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41591-019-0503-6>.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Correspondence and requests for materials** should be addressed to R.K.

**Peer review information:** Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

## Methods

**Data.** Anonymized clinical records of urine culture tests (culture reports) and records of antibiotic purchases (purchase reports) were obtained from MHS from July 2007 to June 2017. Randomly generated patient identifiers were used to link culture reports and antibiotic purchase reports.

**Culture reports.** Antibiotic resistance profiling of bacterial pathogens isolated from urine cultures was carried out centrally (in two locations until 2010 and in one central lab from 2010). We retrieved 711,099 culture reports of positive samples from 315,047 patients total (positive samples indicated bacteriuria, and as samples were most often sent for patients presenting symptoms, we considered these samples as representing UTIs). Each report included the following. (1) Unique patient code. (2) Date of sample. (3) List of isolates cultured with species identification (typically one isolate per sample; 3.6% of samples had more than one isolate). (4) Resistance profile of the isolates from processed results of a VITEK 2 system given as ‘sensitive’, ‘intermediate’ and ‘resistant’ for each drug tested. We focused on resistance to the six antibiotics most commonly prescribed in empirical treatment of these UTIs, with empiric prescription defined as a prescription on the same day the sample was taken, excluding any chance of the measurements being available. ( $N_{\text{Resistances}} = 6$ , Supplementary Table 1 and Table 1; ofloxacin resistance was excluded as measurements were not available as of 2013). Resistance to these antibiotics was routinely measured across the 10-year period, except for cephalexin that was only measured from 2014 (Extended Data Fig. 1). (5) Demographics: age, gender, pregnancy of the patient, as well as an identifier of patients residing in retirement homes.

**Antibiotic purchase reports.** All drug purchases by prescription are routinely recorded in MHS databases. We identified and retrieved all purchases made by patients with culture reports by converting internal MHS drug codes to Anatomical Therapeutic Chemical (ATC) classifications of antibiotics (Supplementary Table 1). Each purchase record included: (1) unique patient code to be linked to the code of the culture record; (2) internal MHS product code, which was translated to an ATC drug code; and (3) date of purchase.

**Choice of drugs for analysis.** We focused on the 11 antibiotic compounds ( $N_{\text{ATC}} = 11$ ), most purchased in the data set (Supplementary Table 1).

**Feature definition.** For each urine sample  $m$ , we define the following parameters used for the logistic regression and the GBDTs.

**Sample resistance profile.** For each urine sample  $m$ , we defined  $Y_k^m$  as 0 for sensitive and intermediate and 1 for resistant to the antibiotic  $k$  ( $1 \leq k \leq N_{\text{Resistances}}$ ). If the sample had multiple isolates,  $Y_k^m$  was assigned as 1 if at least one isolate was resistant. Missing resistance measurements are defined as not available (N/A), and for each antibiotic  $k$  only samples that have defined resistance to it are used when training or testing its logistic regression or GBDT.

**Demographics.**  $X_m^{\text{Gender}}$ , 0/1 for males/females;  $X_m^{\text{Pregnancy}}$ , 0/1 indicating pregnancy;  $X_m^{\text{Ret.Home}}$ , 0/1 indicating residence in retirement homes;  $X_{m,j}^{\text{Age}}$ , 0/1 indicating patient age at the time of UTI sampling in group  $j = 1, 2, \dots, 10$  standing for 0–10, 11–20, ..., 91–100 years;  $X_m^{\text{Date}}$ , date of sample in units of annual quarters starting 2007;  $X_{m,j}^{\text{Season}}$ , 0/1 indicating the quarter of the sample within the calendar year, with  $j = 1, 2, 3, 4$ .

**Sample history.** For a given sample, we consider all earlier samples of the same patient (if any). We assigned the time difference between any such earlier sample and the current sample,  $t = t_{\text{Past sample}} - t_{\text{Sample}}$  ( $t$  is negative, designating past events), into 1 of 16 time bins ( $i = 1, 2, \dots, 16$ ). A bin  $i$  was defined by  $t_i \leq t < t_{i-1}$ , with  $\{t_0, \dots, t_{16}\} = \{-1, 2, 4, 8, 16, 24, 32, \dots, 112\}$  weeks. Boundary choice in integer number of weeks is important to avoid effects of weekends and of patient preference for a specific week day. Previous samples within 1 week of the current sample were not included as they likely represent data on the same infection that might not have been available yet to the physician at the time of the second sample. We then calculated  $X_{m,i,k}^{\text{Previous Resist}}$  and  $X_{m,i,k}^{\text{Previous Sensitive}}$  as the number of prior cultures within time bin  $i$ , whose resistance  $Y_k^m$  equals 1 or 0 (resistant or sensitive), respectively.

**Drug purchase history.** For each urine sample, we considered all earlier drug purchases made by the same patient. We assigned the time difference between the urine sample date and a given past purchase,  $t_{\text{Purchase}} - t_{\text{Sample}}$ , into eight logarithmically spaced time bins ( $i = 1, 2, \dots, 8$ ). A bin  $i$  was defined by  $t_i \leq t < t_{i-1}$ , where the boundaries of these time bins were  $\{t_0, \dots, t_8\} = \{-1, 2, 4, 8, \dots, 128\}$  weeks (the logarithmic binning was chosen to increase statistical power at large time differences where purchase density is lower). For each sample, we then calculated  $X_{m,i,j}^{\text{ATC}}$  as the number of purchases of a given drug  $j$  ( $1 \leq j \leq N_{\text{ATC}}$ ; Supplementary Table 1) made by the patient during time bin  $i$ . For the distribution of purchases per these logarithmically spaced bins, see Supplementary Fig. 5.

**Cross-resistance.** To resolve direct versus indirect associations of drug purchase and resistance, we adjusted the logistic regression of resistance to a given antibiotic  $k$  as a function of past drug purchases by the resistances to all other drugs  $j$  that are non-analogous to  $k$ . We defined  $A_{k,j}$  as a binary variable equal to 0 and 1 for analogous versus non-analogous drug pairs, respectively. ‘Analogous’ pairs were defined as antibiotics that have exceptionally high cross-resistance ( $A_{k,j} = 0$  for  $\text{corr}(Y_k^m, Y_j^m) > A_{\text{threshold}}$ ; we used  $A_{\text{threshold}} = 0.7$  that corresponded to drug pairs of the same class; see pairs labeled with ‘x’ in Extended Data Fig. 5). We then added as features for each sample  $m$  in the regression analysis of a given antibiotic  $k$  the resistance measurements  $Y_j^m$  to all antibiotics  $j$  for which  $A_{k,j} = 1$ . Note that these cross-resistance features provided information from the focal sample and were used only in the analysis of direct/indirect effect of purchases (Fig. 4b) and not for the evaluation of resistance predictability.

**Logistic regression.** Logistic regression of resistance for each antibiotic was performed via the Matlab glmfit function. For each of the resistances  $k = 1, 2, \dots, 6$ , the probability of resistance  $P_k$  was fitted to the sample resistance  $Y_{m,k}$  for all urine samples that had measurement of resistance to  $k$  either across the entire 10-year data set (for Figs. 2 and 4) or across the ‘training period’ (for the analysis of the predictive power of Fig. 5; see Extended Data Fig. 1 for a definition of the training period for each of the six antibiotics). The different logistic models included combinations of the following ten terms:

	Term
$\ln\left(\frac{P_k^m}{1-P_k^m}\right) = C_k^{\text{Const}} +$	
$C_k^{\text{Gender}} X_m^{\text{Gender}} +$	1
$C_k^{\text{Pregnancy}} X_m^{\text{Pregnancy}} +$	2
$C_k^{\text{Ret.Home}} X_m^{\text{Ret.Home}} +$	3
$\sum_{j=2}^{10} C_{k,j}^{\text{Age}} X_{m,j}^{\text{Age}} +$	4
$C_k^{\text{Date}_1} \left(\frac{X_m^{\text{Date}}}{4}\right) + C_k^{\text{Date}_2} \left(\frac{X_m^{\text{Date}}}{4}\right)^2 +$	5
$\sum_{j=1}^3 C_{k,j}^{\text{Season}} X_{m,j}^{\text{Season}} +$	6
$\sum_{i=1}^{16} C_{k,i}^{\text{Previous Resist}} X_{m,k,i}^{\text{Previous Resist}} + C_{k,i}^{\text{Previous Sensitive}}$	7
$X_{m,k,i}^{\text{Previous Sensitive}} +$	
$\sum_{i=1}^8 \sum_{j=1}^{N_{\text{ATC}}} C_{k,i,j}^{\text{ATC}} X_{m,i,j}^{\text{ATC}} +$	8
$\sum_{j=8}^{10} C_{k,j}^{\text{Age} \times \text{Ret.Home}} X_{m,j}^{\text{Age}} X_{m,j}^{\text{Ret.Home}} +$	9
$\sum_{j=1}^{N_{\text{Resistances}}} A_{k,j} C_{k,j}^{\text{Cross Resist}} Y_j^m$	10

Different combinations of the above terms were used in the different regression models as shown in Supplementary Table 3 (each row in the table represents a logistic model that was applied to each of the six antibiotics).

**Calculating odds ratios from logistic regression.** For each antibiotic  $k$ , ORs were calculated from the coefficients of above logistic regressions.

**Binary variables.** For the binary variables gender, pregnancy and retirement home, ORs were defined as:  $\text{OR}_k^{\text{Gender}} = \exp(C_k^{\text{Gender}})$  female versus male,  $\text{OR}_k^{\text{Pregnancy}} = \exp(C_k^{\text{Pregnancy}})$  pregnant versus non-pregnant,  $\text{OR}_k^{\text{Ret.Home}} = \exp(C_k^{\text{Ret.Home}})$  retirement home residence versus patients not residing in retirement homes.

**Categorical variables.** For the categorical variables age and season, ORs for each category relative to the reference (age group of 0–10 years, fourth quarter, respectively) is given by  $\text{OR}_{k,j}^{\text{Age}} = \exp(C_{k,j}^{\text{Age}})$  and  $\text{OR}_{k,j}^{\text{Season}} = \exp(C_{k,j}^{\text{Season}})$ , where  $C_{k,j}^{\text{Age}}$  and  $C_{k,j}^{\text{Season}}$  are reported in Supplementary Table 2. In Fig. 2, we report for age  $\text{OR}_{k,\text{Age max}}^{\text{Age}} = \exp(C_{k,j_{\text{max}}}^{\text{Age}})$ , with  $j_{\text{max}} = 10$  standing for the 91–100 year group; and for season,  $\text{OR}_{k,\text{Season max}}^{\text{Season}} = \exp(C_{k,j_{\text{max}}}^{\text{Season}})$ , with  $j_{\text{max}} = 2$  standing for the second quarter (most contrast to the reference, which is the fourth quarter).

**Quadratic variables.** For date, which is fitted quadratically, the individual regression coefficients and their CIs are reported in Supplementary Table 2. In Fig. 2b, we also report, for each antibiotic  $k$ , effective ORs defined as the ratios between the maximal and minimal expected odds taken across the relevant date range of ( $0 \leq X^{\text{Date}} \leq 40$ ):

$$\text{OR}_k^{\text{Date}} = \exp[\max_{0 \leq x \leq 1} (C_k^{\text{Date}_1} x + C_k^{\text{Date}_2} x^2) - \min_{0 \leq x \leq 1} (C_k^{\text{Date}_1} x + C_k^{\text{Date}_2} x^2)]$$

Note that when these quadratic dependencies are monotonic within the relevant range ( $0 \leq x \leq 1$ ), the above formula becomes simply  $\text{OR}_k^{\text{Date}} = \exp(C_k^{\text{Date}_1} + C_k^{\text{Date}_2})$ .

**Analysis of ‘memory’ across sample pairs.** To analyze ‘memory’ of resistance across samples, we considered all pairs of samples from the same patient (across all patients with 2–10 samples) and binned them according to their time difference  $t = t_1 - t_2$  (where  $t_1$  and  $t_2$  are the sample dates of the early and late sample;  $t$  is always negative, indicating information on current sample from past samples) into time bins as indicated by the bars in Fig. 3. In each time bin and for each antibiotic, we counted  $N_{R \rightarrow R}$ ,  $N_{R \rightarrow S}$ ,  $N_{S \rightarrow R}$  and  $N_{S \rightarrow S}$  as the number of urine sample pairs where the early and late samples were resistant, or sensitive (for example  $N_{R \rightarrow S}$  was the number of same-patient sample pairs, within the time difference bin, where first sample was resistant and the second sensitive to the given focal antibiotic. For each antibiotic, only samples for which resistance was measured were considered). We then calculated for each time difference bin the risk ratio  $\zeta_{\text{pairs}} = [N_{R \rightarrow R}/(N_{R \rightarrow R} + N_{R \rightarrow S})]/[N_{S \rightarrow R}/(N_{S \rightarrow R} + N_{S \rightarrow S})]$ .

**Gradient-boosting decision trees.** GBDTs are an ensemble method combining regression trees with weak individual predictive performances, into a single high-performance model. This is done by iteratively fitting decision trees with each iteration targeting the prediction residuals of the preceding tree. The final model is built by combining weighted individual tree contributions, with weights proportional to their performances. For each of the six antibiotics, a boosted decision tree ensemble was fitted using all features as defined above (demographics, sample history and drug purchase history) on the training set as defined by the training time period (Extended Data Fig. 1, green bars). This training data set was sampled to balance resistant/sensitive label frequency. For parameter tuning, a validation data set was sampled from the training set to be used for model selection (20%). For the estimator of the  $i$ th iteration, a decreasing learning rate  $\eta_i$  was used such that  $\eta_i = \eta_0 \alpha^i$ , with an annealing rate  $\alpha = 0.99$  and an initial learning rate  $\eta_0 = 0.1$ . To further promote a diverse ensemble of individual estimators, a 0.9 feature-sampling and observation-sampling rates were used. Fitting of interaction effects is controlled by varying the size of the individual regression trees, with tree estimator of depth  $k$  producing models with up to  $k$ -way interactions. The model was tuned to match data complexity by iteratively increasing tree depth limit of all ensemble estimators while evaluating performance on the validation set, selecting the best depth for each antibiotic.

**Unconstrained algorithm for drug choice.** Given the complete-model machine-learning assigned probabilities of resistance  $P_k^m$  of each same-day empirically treated infection  $m = 1, 2, \dots, N_{\text{samples}}$  to each of the antibiotics  $k = 1, \dots, N_{\text{Resistances}}$ , the unconstrained model simply recommends for each infection, the antibiotic  $K_{\text{rec}}^m$  for which the model predicted probability of resistance is lowest. Namely,  $K_{\text{rec}}^m$  is defined by  $P_{K_{\text{rec}}^m}^m = \min_k(P_k^m)$ .

**Constrained (cost-adjusted) algorithm for drug choice.** The constrained, cost-adjusted, algorithm for drug choice takes as input the complete-model machine-learning assigned probabilities of resistance  $P_k^m$  of each same-day empirically treated infection  $m = 1, 2, \dots, N_{\text{samples}}$  to each of the antibiotics  $= 1, \dots, N_{\text{Resistances}}$ , as well as the target total number of uses of each drug  $n_k^{\text{target}}$  (with  $\sum_{k=1}^{N_{\text{Resistances}}} n_k^{\text{target}} = N_{\text{samples}}$ ). The algorithm needs to return as output the optimal recommended drug treatments  $K_{\text{rec}}^m$  for each infection  $m$  such that the overall expected rate of mismatched treatment  $\sum_{m=1}^{N_{\text{samples}}} P_{K_{\text{rec}}^m}^m$  is minimized while the overall usage of each drug  $n_k = \sum_{m=1}^{N_{\text{samples}}} \delta(k, K_{\text{rec}}^m)$  (where  $\delta(i, j) = 1$  for  $i = j$  and 0 otherwise) satisfies  $n_k = n_k^{\text{target}}$  for all the antibiotics  $k$ . This constrained optimization problem can be solved exactly. First, we adjust the machine-learning model probabilities of resistance to each antibiotic by an additive drug-specific value  $C_k$  accounting for an assigned ‘cost’ of using this drug:  $Q_k^m = P_k^m + C_k$ . Then, given a set of cost values for all the antibiotics  $\{C_k\}$ , the recommended antibiotic  $K_{\text{rec}}^m$  for each infection  $m$  is defined by  $Q_{K_{\text{rec}}^m}^m = \min_k(Q_k^m)$  and given these drug choice  $K_{\text{rec}}^m$  for all the infections, we then calculate the overall drug distribution  $n_k = \sum_{m=1}^{N_{\text{samples}}} \delta(k, K_{\text{rec}}^m)$ . These drug distribution counts are therefore a function of the cost values  $n_k = n_k(\{C_k\})$ . We then numerically solve for the set of cost values  $\{C_k^{\text{target}}\}$  for which the drug distribution satisfies  $n_k(\{C_k^{\text{target}}\}) = n_k^{\text{target}}$ . For  $N_{\text{Resistances}} = 6$ , this amounts to numerically solving six equations with the six  $C_k$ s as variables (the degeneracy due to  $\sum_{k=1}^{N_{\text{Resistances}}} n_k^{\text{target}} = N_{\text{samples}}$  is offset by an added normalization  $\sum_k C_k = 0$ ). Once we solved for the cost values  $\{C_k^{\text{target}}\}$ , the specific drug recommendations  $K_{\text{rec}}^m$  for each infection were defined by  $Q_{K_{\text{rec}}^m}^m = \min_k(Q_k^m)$  with  $Q_k^m = P_k^m + C_k^{\text{target}}$ .

It is easy to prove mathematically that this solution optimally minimizes risk of resistance given the constraints of the total usage of each drug. Let us assume that there exists an alternative solution  $K_{\text{alt}}^m$  that has the same distribution of drug usage but with a lower predicted chance of resistance  $\sum_{m=1}^{N_{\text{samples}}} P_{K_{\text{alt}}^m}^m < \sum_{m=1}^{N_{\text{samples}}} P_{K_{\text{rec}}^m}^m$ . As the two solutions have the same overall number of uses of each drug, there must exist a set of pairwise swapping steps that transforms the ‘rec’ solution to the ‘alt’ solution, where each step consists of taking two infections  $m_1$  and  $m_2$  and swapping

their recommended prescriptions  $K_{\text{rec}}^{m_1}$  and  $K_{\text{rec}}^{m_2}$  (an operation that maintains the same overall use of the drugs). But, given that the recommended prescriptions  $K_{\text{rec}}^{m_1}$  and  $K_{\text{rec}}^{m_2}$  are defined by  $Q_{K_{\text{rec}}^{m_1}}^{m_1} = \min_k(Q_k^{m_1})$  and  $Q_{K_{\text{rec}}^{m_2}}^{m_2} = \min_k(Q_k^{m_2})$ , swapping them necessarily leads to equal or higher overall probability of mismatched treatment:

$$\begin{aligned} P_{K_{\text{rec}}^{m_1}}^{m_1} + P_{K_{\text{rec}}^{m_2}}^{m_2} &= Q_{K_{\text{rec}}^{m_1}}^{m_1} - C_{K_{\text{rec}}^{m_1}}^{\text{target}} + Q_{K_{\text{rec}}^{m_2}}^{m_2} - C_{K_{\text{rec}}^{m_2}}^{\text{target}} \geq \\ Q_{K_{\text{rec}}^{m_1}}^{m_1} - C_{K_{\text{rec}}^{m_1}}^{\text{target}} + Q_{K_{\text{rec}}^{m_2}}^{m_2} - C_{K_{\text{rec}}^{m_2}}^{\text{target}} &= P_{K_{\text{rec}}^{m_1}}^{m_1} + P_{K_{\text{rec}}^{m_2}}^{m_2} \end{aligned}$$

Therefore, any swap among the set of infections of the drugs recommended by the algorithm leads to an increased predicted rate of mismatched treatment. The solution we provide is therefore optimal.

Finally, we note that an important added value of this approach is that it also provides the cost values  $\{C_k^{\text{target}}\}$  for each of the antibiotics. Namely, given the distribution of antibiotics prescribed by physicians, we can deduce effective cost values that effectively account for the different global considerations physicians take such as ease of use, and tendency to avoid drugs of last resort. Once these cost values are determined, such as based on the 1-year test period, they can be used for future algorithmic recommendations of drug prescriptions. Namely, for a given new case with machine-learning probability of resistance  $P_k$  for each of the antibiotics  $k$ , the algorithm will simply recommend the antibiotic  $K_{\text{rec}}$  for which  $Q_{K_{\text{rec}}} = \min_k(Q_k)$ , where  $Q_k = P_k + C_k^{\text{target}}$ .

**Analysis of same-day empirical treatments.** We identified all cases across the 1-year test period where patients purchased one (and only one) of the six antibiotics on the same day they had a sample sent for culture and for which resistances to all six antibiotics were measured (SDET). We then retrospectively annotated each SDET prescription as ‘matched’ or ‘unmatched’ according to whether the sample was sensitive or resistant to the prescribed antibiotic, respectively. The rate of mismatched treatment was then defined across all of these SDET patients (Fig. 5c), as well as separately across all of the patients treated with a given drug (Fig. 5d, top). A similar analysis was done for the drugs recommended by either the unconstrained or the constrained (cost-adjusted) models (Fig. 5c,d). Mismatch rates were also compared with two models of null expectations. In the ‘dice’ model, we randomly chose, for each SDET case, one of the six drugs with equal probability. In the ‘Random permutation’ model, we randomly permuted across the SDET cases the same overall pool of drugs prescribed by the physicians (thereby maintaining the exact same frequency of use of each of the six drugs). For each of these models, we repeated 1,000 random simulations and calculated the average mismatched treatment rate (Fig. 5c, horizontal lines).

**Statistical significance of mismatched treatment rates.** We performed 10,000 bootstrapping simulations in which we randomly sampled, with replacement, 11,952 cases from the 11,952 SDET cases and calculated for each of these 10,000 simulations the mismatch rate for the prescriptions given by physicians, the constrained machine-learning model (CML), the unconstrained machine-learning model (UCML), the random permutation model (RP) and the random dice model (RD). For each of these five models, we reported the 95% CI of the mismatched treatment rate based on the 2.5th and 97.5th percentile values of the mismatched treatment rate of the specified model across the 10,000 bootstrapping simulations. When comparing two models, we consider the difference between the mismatched treatment rates of the two models for each of the 10,000 simulations. For all reported model comparisons (physicians–RD, physicians–RP, UCML–physicians and CML–physicians), the mismatch rate in the first model was lower than the mismatch rate in the second model in virtually all 10,000 bootstrapping simulations (representing  $P$  values lower than  $10^{-4}$ ). As an estimate for the  $P$  value, we reported the error function based on the average and standard deviation of the difference of mismatch rate between the two models across the 10,000 bootstrapping simulations.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

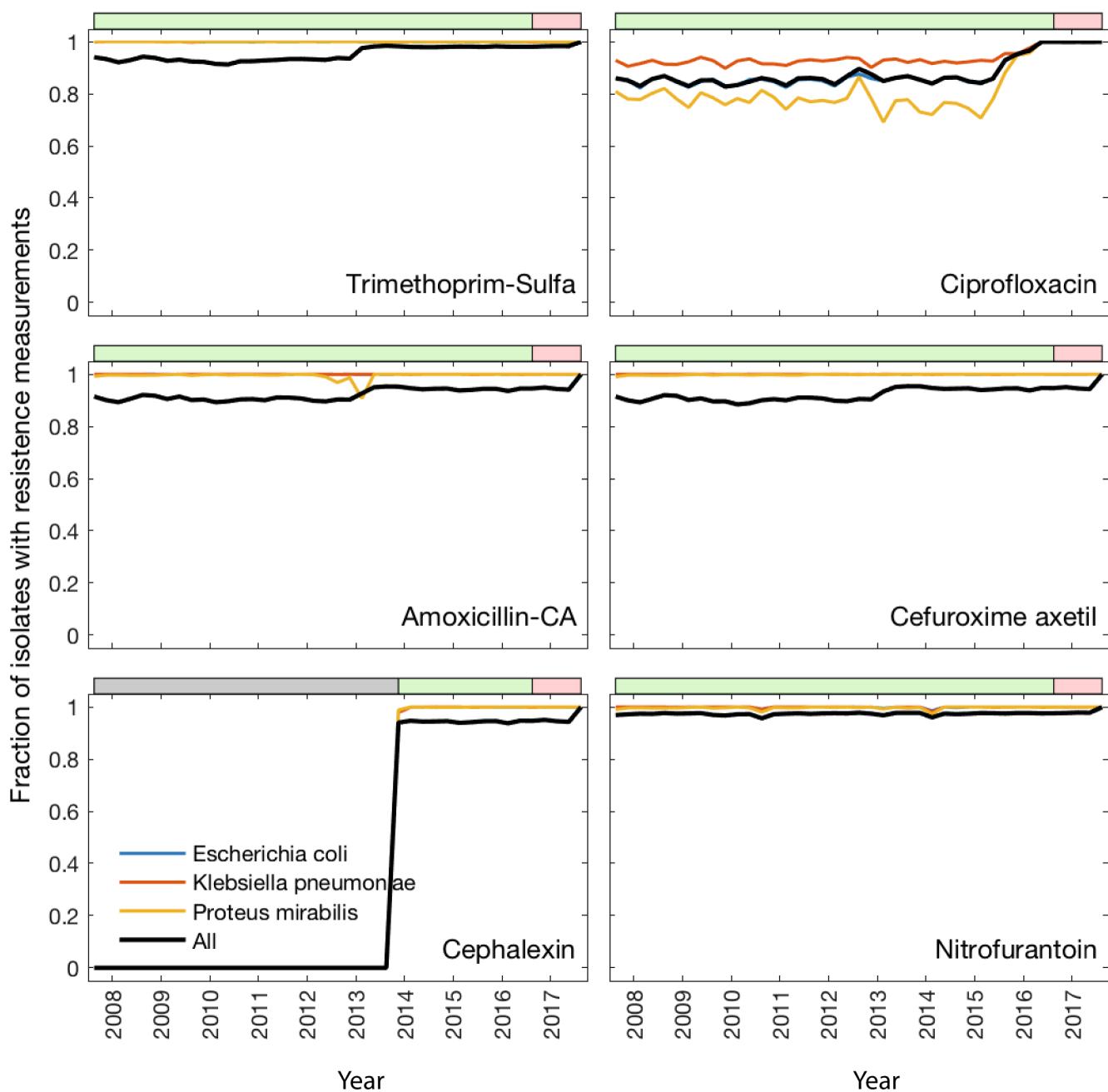
**Ethical approval.** The study protocol was approved by the ethics committee of Assuta Medical Center, Tel-Aviv, Israel.

## Data availability

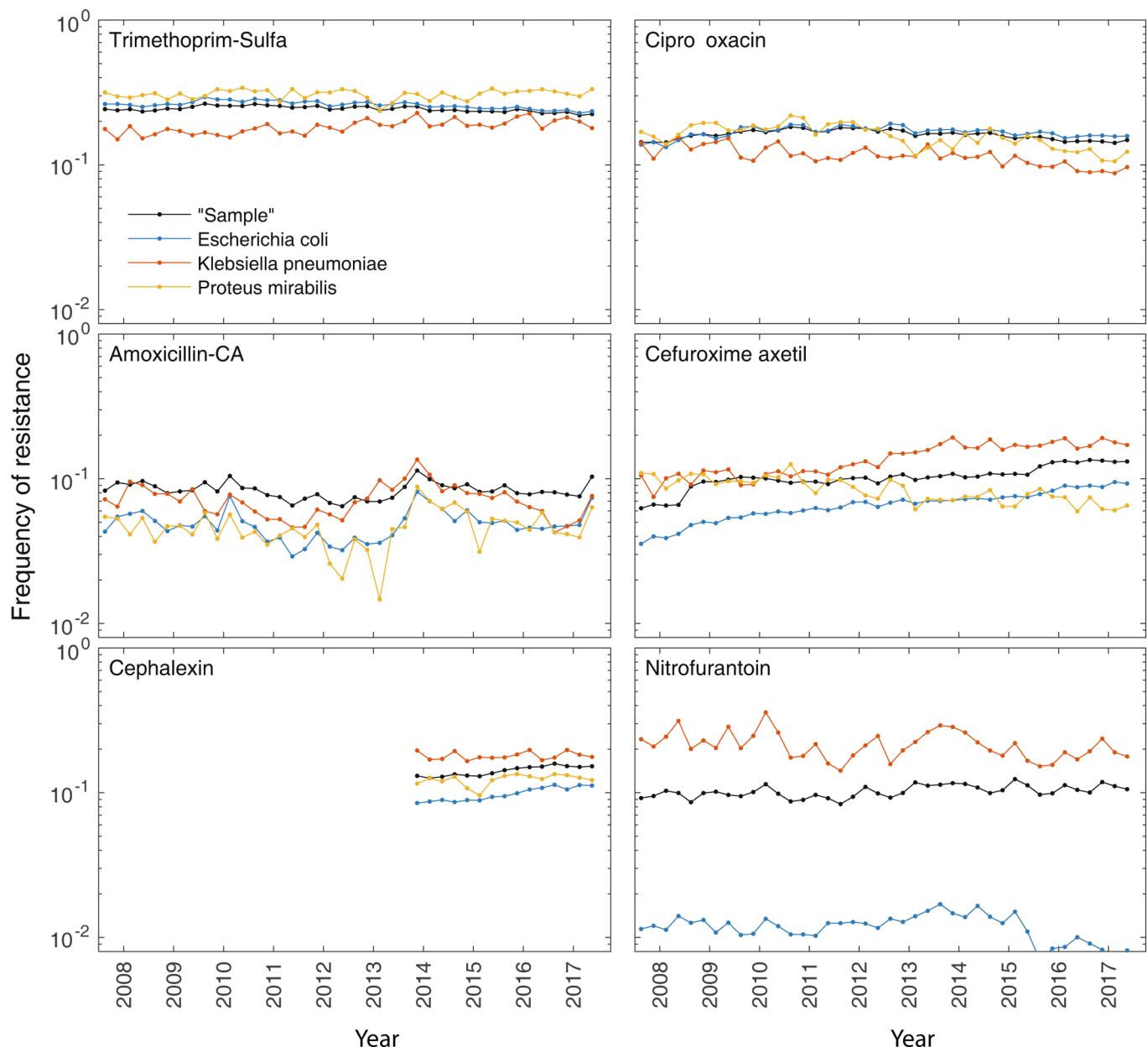
The data that support the findings of this study are available from Maccabi Healthcare Services but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Access to the data is, however, available upon reasonable request and signing an MTA agreement with Maccabi Healthcare Services.

## Code availability

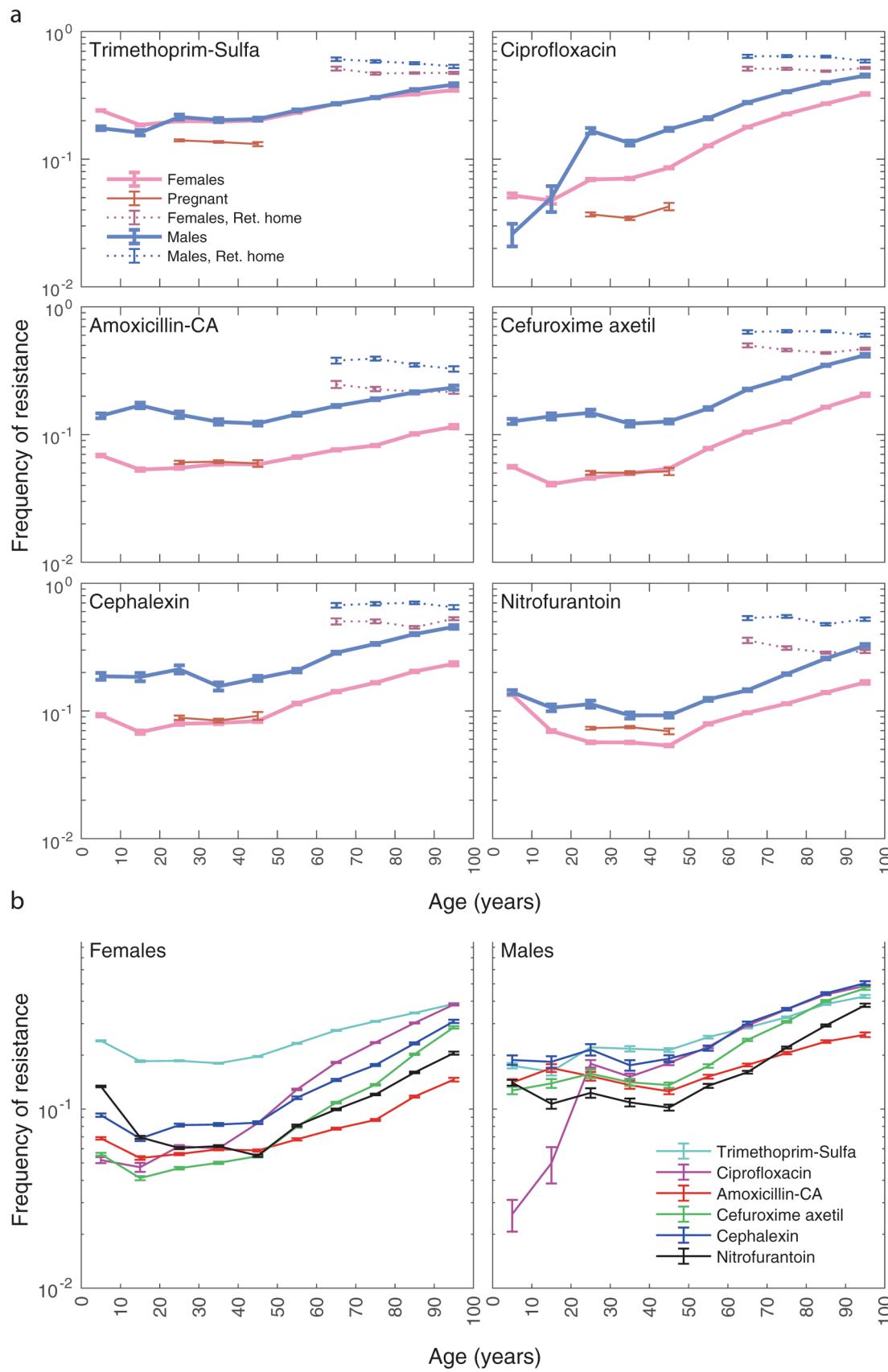
Code used for data analysis is available upon request.



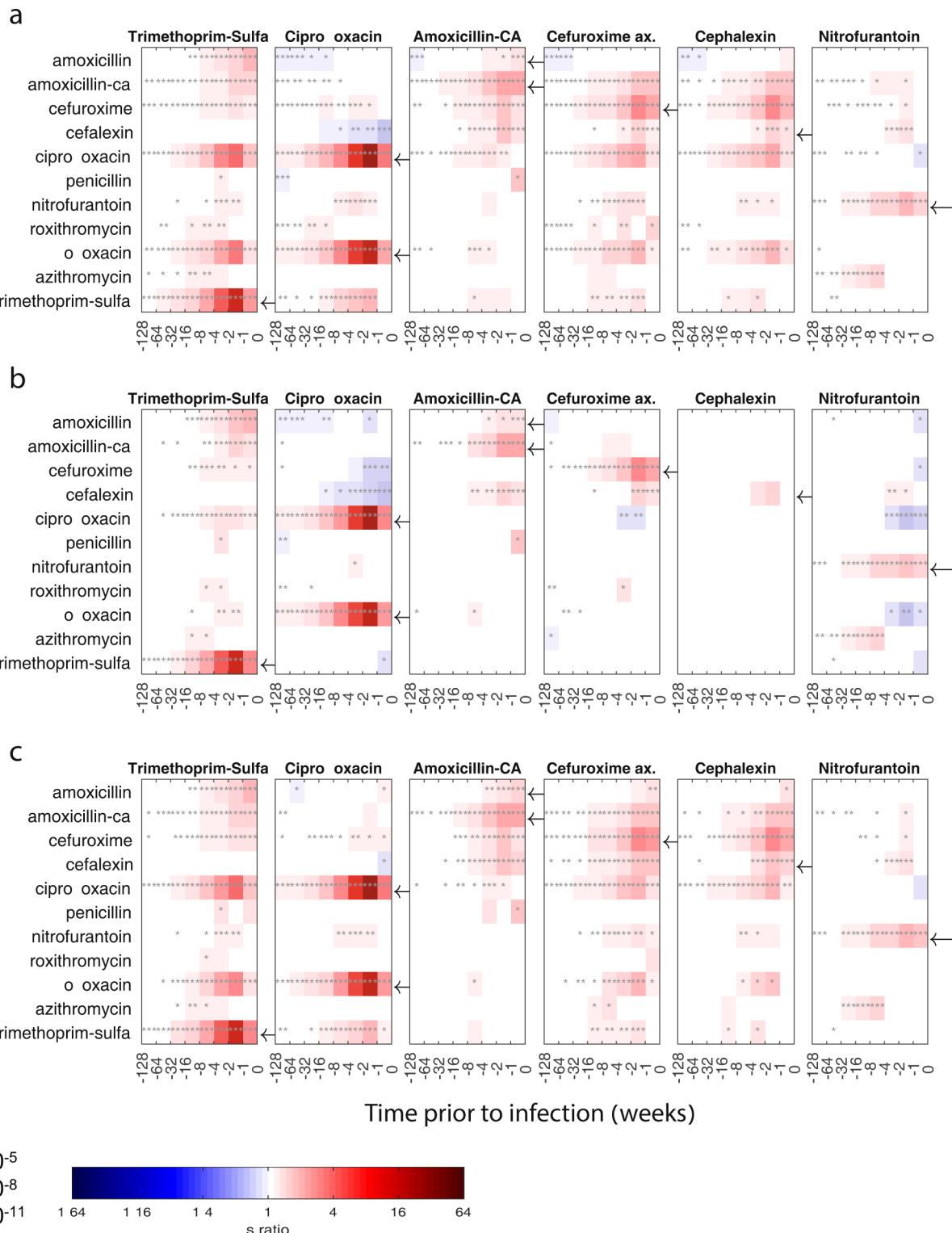
**Extended Data Fig. 1 | Availability of resistance measurements over time.** For each of the six antibiotics, the fraction of urine samples for which resistance was measured, overall (black) and for each of the three most common species (colors), is plotted across the 10-year sampling period. Also indicated are the time ranges used for model training (green horizontal bars) and testing (red bars). Time periods during which measurements of resistance to cephalexin were scarce and were removed from analysis (gray bar).



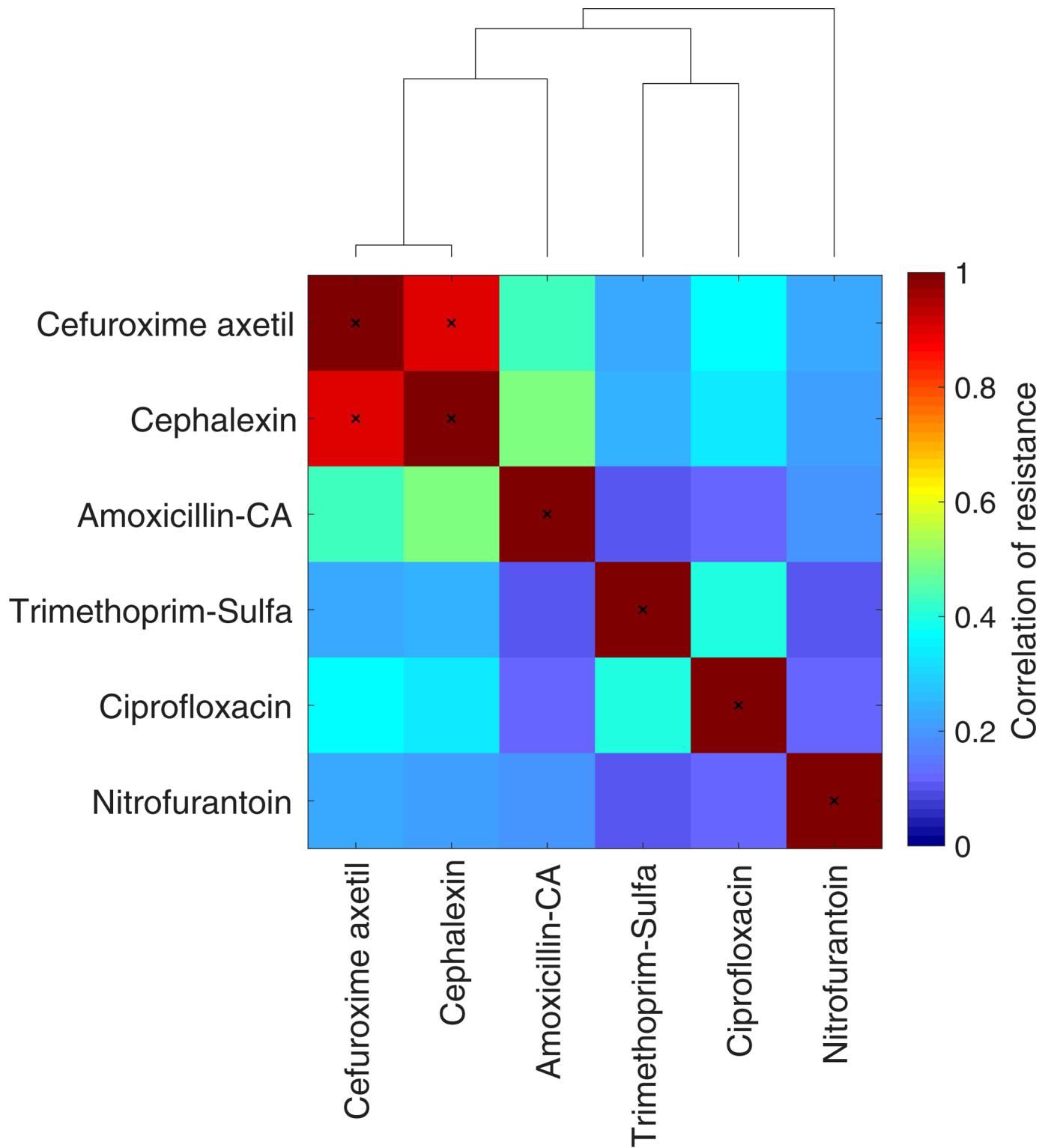
**Extended Data Fig. 2 | Frequency of resistance over time.** Frequencies of resistance for each of the three common species (colored lines) and the overall sample (black lines) over the 10-year data set. Empty time intervals correspond to periods during which resistance was not frequently measured (matching the gray horizontal bar of Extended Data Fig. 1).



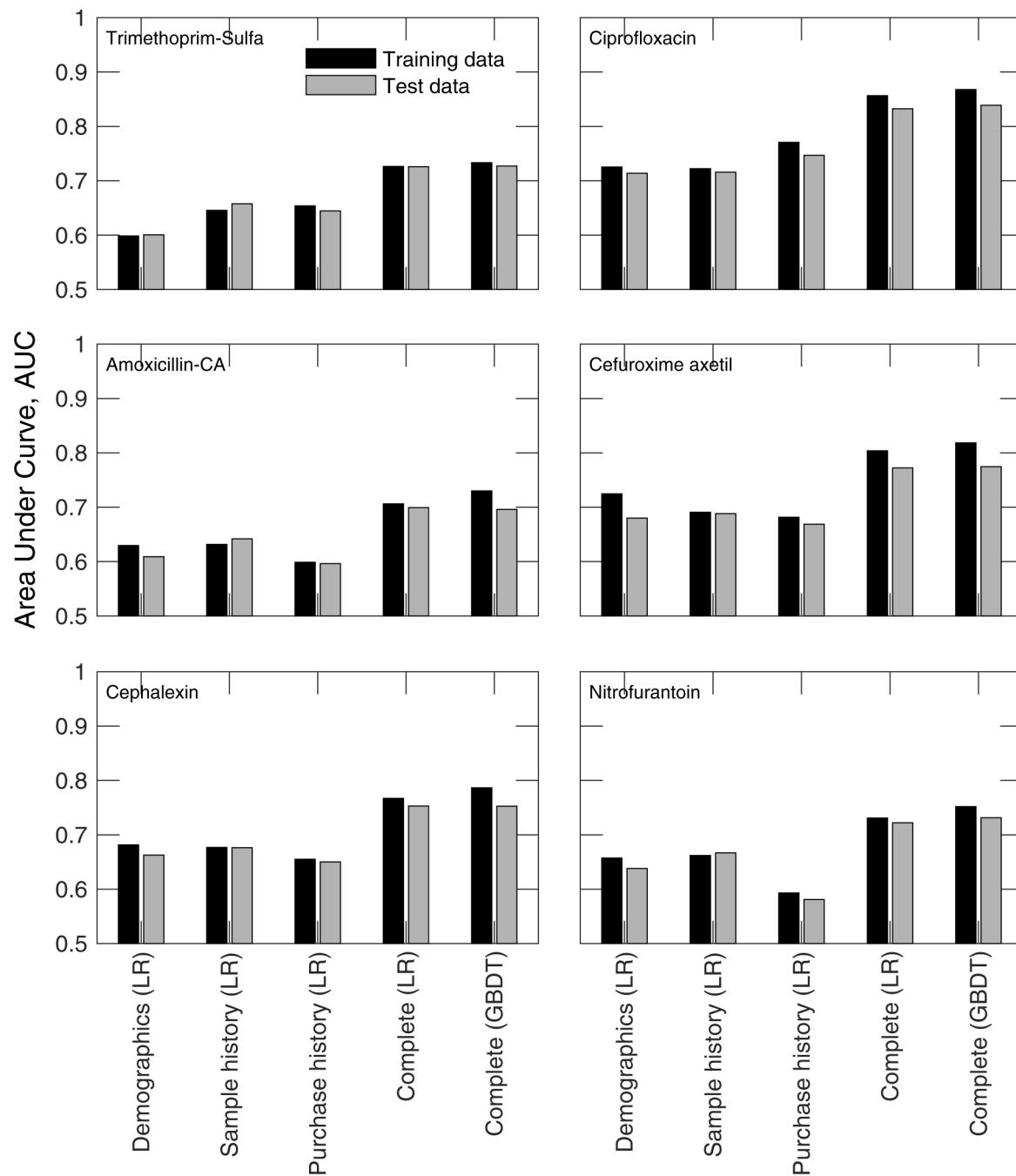
**Extended Data Fig. 3 | Frequency of resistance as a function of age for different demographic groups.** Frequency of resistance to each of the six antibiotics, in each of ten age bins ( $0, 10, \dots, 100$  years). **a**, Frequencies of resistance for five non-overlapping demographic groups: men not residing in retirement homes (blue), men residing in retirement homes (dotted blue), women not pregnant and not residing in retirement homes (magenta), women in retirement homes (magenta dotted) and pregnant women (red). **b**, Comparing the overall frequency of resistance to the six drugs for women and men across age.



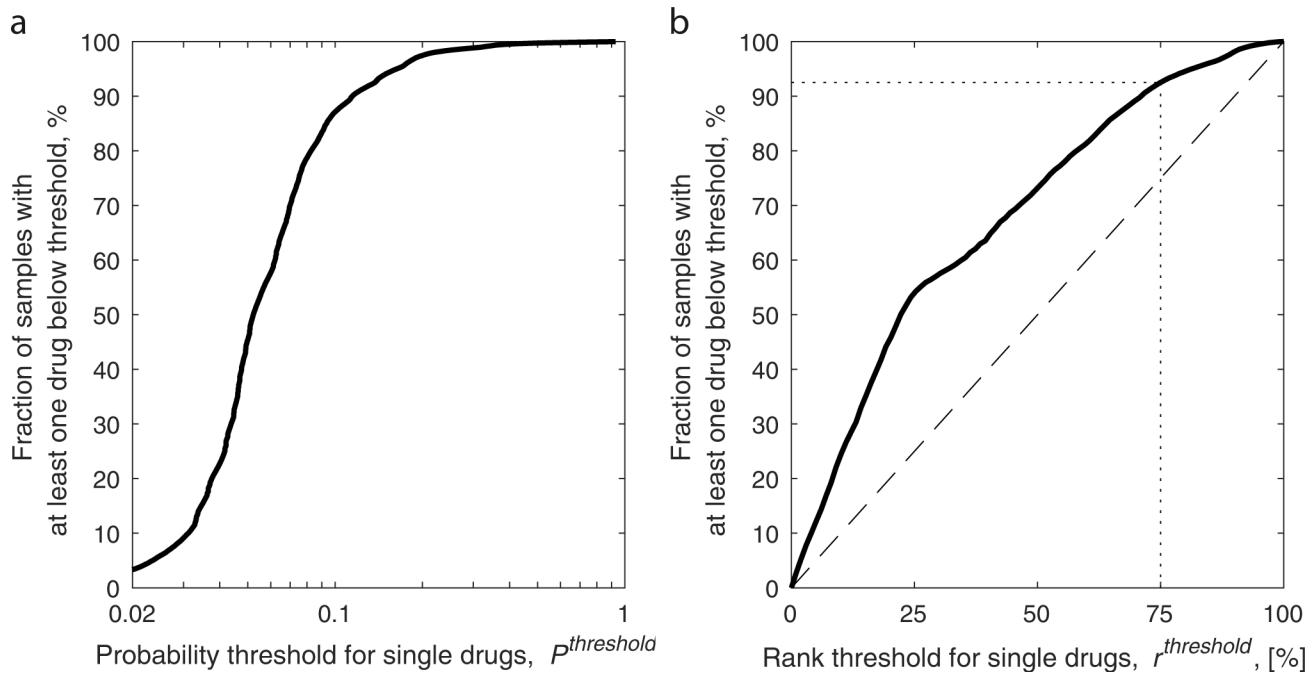
**Extended Data Fig. 4 | Odds ratios of resistance to each of the antibiotics for past purchases of different drugs across a range of purchase-to-sample time intervals: adjustments for demographics and cross-resistance.** **a**, Multivariate logistic regression models for the association of each antibiotic resistance with past purchases of the indicated drugs not accounting for cross-resistance (Methods: Logistic regression ‘Purchase history’. Same graphical scheme as in Fig. 4a,b). **b**, Logistic regression model as in **a** adjusted for cross-resistance (Methods: Logistic regression ‘Purchase history adjusted for cross-resistance’). **c**, Logistic regression model as in **a** adjusted for demographics (Methods: Logistic regression ‘Purchase history adjusted for demographics’). Gray asterisks indicate statistical significance and non-significant values, with Bonferroni corrected  $P > 0.05$ , are blanked.



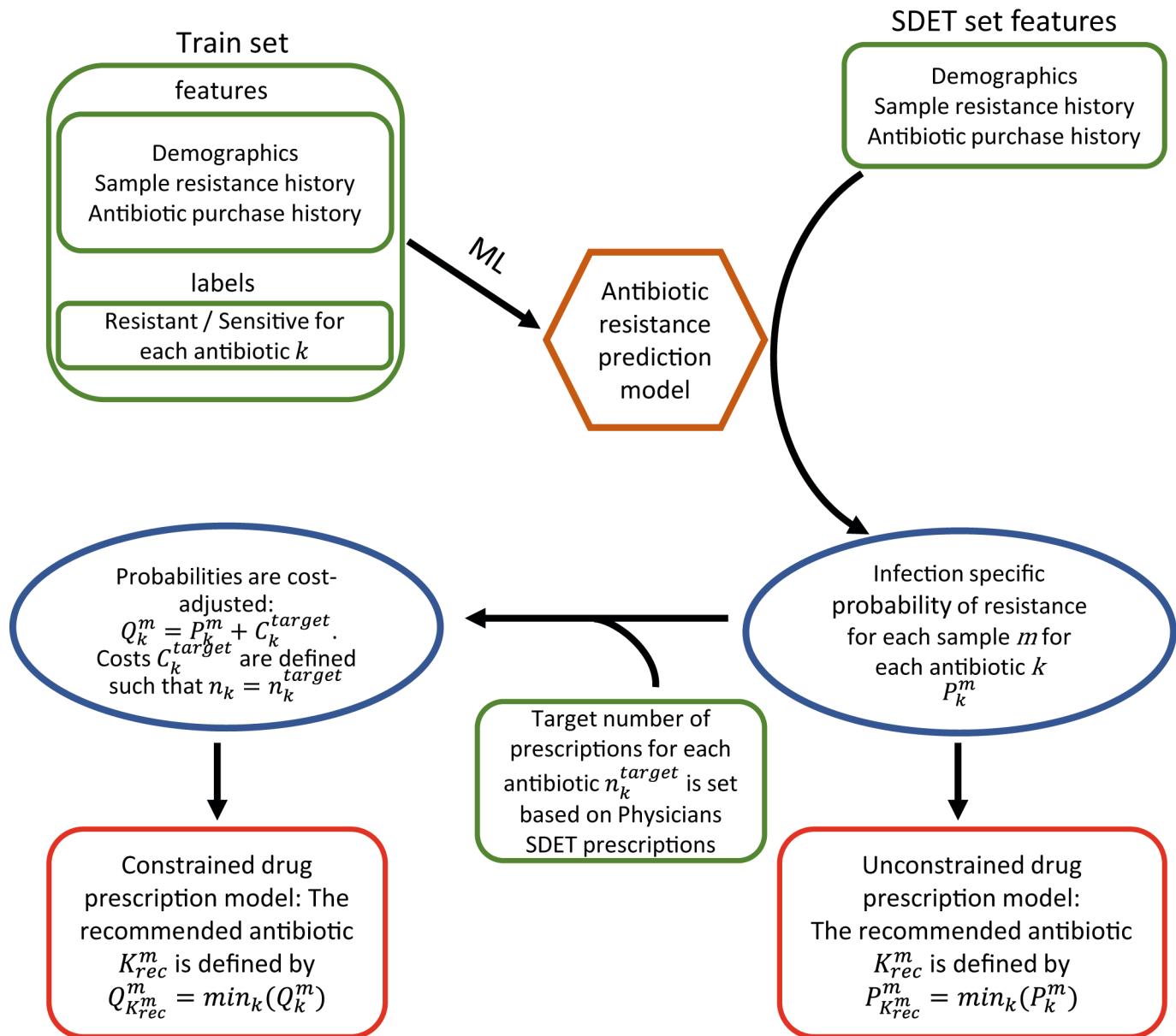
**Extended Data Fig. 5 | Correlations among resistances to different antibiotics.** Correlation among resistance measurements for each pair of antibiotics across all samples for which both resistances were measured. Cephalexin and cefuroxime axetil, which have a particularly high correlation (marked with 'x'), were treated as 'analogous' in the analysis of indirect effects of purchases on resistance (Methods: Logistic regression 'Purchase history adjusted for cross-resistance').



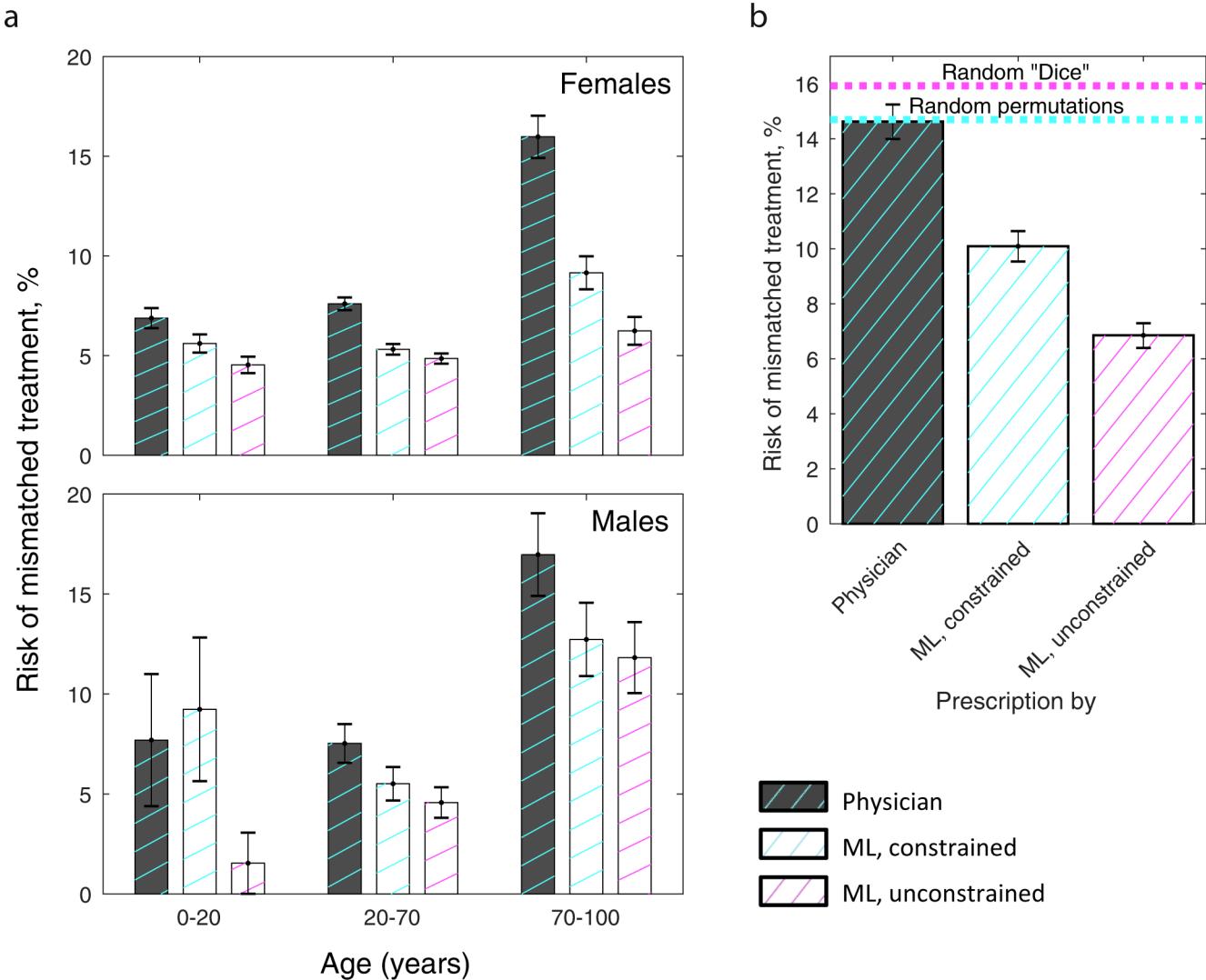
**Extended Data Fig. 6 | Model performance on test and training data.** AUC for receiver operator characteristic for prediction of resistance based on demographics, sample history and purchase history, individually and in a complete model combining all feature sets. Each feature set was modeled using LR, and the complete model was modeled by both LR and GBDT. To identify overfitting, model performance on the testing data set (gray) was contrasted with model performance on the training data set (black; Extended Data Fig. 1 for definition of training and test time periods). A low level of overfitting is seen for all drugs except trimethoprim, which showed no overfitting. See Supplementary Fig. 4 for feature importance in GBDT models.



**Extended Data Fig. 7 | The fraction of samples that can be treated by at least one drug given set thresholds on the single-drug resistance probability scores.** Given the complete-model-assigned probabilities of resistance  $P_k^m$  of each sample  $m$  to each antibiotic  $k$ , we calculated the fraction of samples, within the 1-year test period, that have at least one drug with resistance score below a threshold. **a,b**, This fraction is calculated assuming that the threshold used to determine resistance of single drugs is either: the same probability threshold  $P^{threshold}$  for all drugs (counting all samples for which  $P_k^m < P^{threshold}$  for at least one antibiotics  $k$ ) (**a**) or the same rank threshold  $r^{threshold}$  for all drugs, counting all samples for which  $P_k^m < P_k^{threshold}$  ( $r^{threshold}$ ) for at least one antibiotics  $k$ , where  $P_k^{threshold}$  ( $r^{threshold}$ ) is the probability threshold of drug  $k$  that includes a fraction  $r^{threshold}$  of the samples (**b**).



**Extended Data Fig. 8 | Schematic diagram of machine-learning-trained prescription models.** A set of samples with features of demographics, sample resistance history and antibiotic purchase history labeled for resistance to each antibiotic  $k$  ('train set') is used to train an antibiotic resistance prediction model (Methods: Logistic regression, terms 1–9). The model is applied to an SDET set of cases from the test period to calculate the probabilities of resistance to each antibiotic. In an unconstrained model the antibiotic with minimal probability for resistance is suggested. The calculated probabilities of resistance together with the respective prescriptions of the SDET set of cases are used to add a 'cost' term. In a constrained drug prescription model, the antibiotic with the minimal cost-adjusted probability is suggested.



**Extended Data Fig. 9 | Robustness of machine-learning-trained prescription models across age and gender and with respect to the clinical definition of resistance.** **a**, Frequency of mismatched treatment across all SDET cases, comparing physician's prescriptions (dark bar) to algorithmic recommendations by the constrained and unconstrained models (cyan and magenta hatched, respectively) for females (top) and males (bottom) separated into three major age groups. **b**, Frequency of mismatched treatment across all SDET cases (Methods), when classifying 'Intermediate' level of resistance as 'resistant'. Comparing mismatch frequencies of physicians' prescriptions (dark bar) to algorithmic recommendations (light bars), either unconstrained (magenta hatched) or constrained for recommending drugs at the same ratio as physicians (cyan hatched). Also presented are the null expectations for randomly prescribing drugs with equal probabilities (random 'dice', magenta dashed) or for random drug permutations (random permutations, cyan dashed).

Corresponding author(s): Roy Kishony

Last updated by author(s): May 19, 2019

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

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

n/a Confirmed

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

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used for data collection.

Data analysis

Data was analysed using custom MATLAB code.

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

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from Maccabi Healthcare Services but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Access to the data is however available upon reasonable request and signing an MTA agreement with Maccabi Healthcare Services.

# Field-specific reporting

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

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

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

Sample size	We analysed the measurement results of 711,099 urine samples collected from 315,047 patients. These is the entire set of urine culture samples measured by Maccabi Healthcare Services during the study period (July 2007 - June 2017).
Data exclusions	We did not exclude any data.
Replication	Data analysis is reproducible.
Randomization	We did not randomize data records.
Blinding	Blinding the investigators was not relevant to the analysis procedures.

## Reporting for specific materials, systems and methods

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

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging