Wastewater surveillance reveals patterns of antibiotic resistance across the United States

Authors: Sooyeol Kim[1], Alessandro Zulli[2], Elana M. G. Chan[2], Dorothea Duong[3], Rebecca Y. Linfield[4], Caroline McCormack[1], Bradley J. White[3], Marlene K. Wolfe[5], Alexandria B. Boehm[2], Amy J. Pickering[1,6,7]

Corresponding author: pickering@berkeley.edu

Affiliation
[1]Department of Civil and Environmental Engineering, University of California, Berkeley, CA, 94720, USA
[2]Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, 94305, USA
[3]Verily Life Sciences, South San Francisco, CA, 94080, USA
[4]Division of Infectious Diseases and Geographic Medicine, Department of Medicine, Stanford University, Stanford, CA, 94305, USA
[5]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, 30329, USA
[6]Chan Zuckerberg Biohub San Francisco, San Francisco, California
[7]Blum Center for Developing Economies, University of California

Abstract
Antibiotic resistance is a growing public health threat, with over 2.8 million antibiotic-resistant infections and 35,000 attributable deaths annually in the U.S. This is an underestimate, as it is based on people who seek medical attention. Here, we sought to use wastewater monitoring to assess community-level antibiotic resistance. This study quantifies concentrations of antibiotic resistance genes (ARGs) by digital droplet PCR in wastewater solids obtained from 163 wastewater treatment plants across the United States. We measure 11 ARGs that confer resistance to beta-lactams (*CMY, CTX-M, KPC, NDM, mecA, OXA-48, TEM, VIM*), colistin (*mcr-1*), tetracycline (*tetW*), and vancomycin (*vanA*). The Northeast and South have higher overall ARG concentrations compared to the West and Midwest. We pair these data with national data sets including antibiotic use, social vulnerability, size of animal agriculture operations, density of healthcare facilities, and presence of airports to investigate potential drivers of resistance. We also generate predictive maps of ARG concentrations for every county in the United States. We show social vulnerability indicators (overcrowding, housing burden, and access to health insurance) and indicators of international travel are associated with increased ARG concentrations in wastewater, while antibiotic usage is only weakly positively correlated. Our results provide a national baseline of ARG concentrations and highlight the complexity of factors driving spread of antibiotic resistance.

Introduction

Antimicrobial resistance (AMR) is a critical and growing global public health threat, with resistant bacterial infections, such as carbapenem resistant *Enterobacterales* and vancomycin resistant *Enterococcus*, becoming more prevalent.[1] The 2019 Antibiotic Resistance Threats Report by the U.S. Centers for Disease Control and Prevention (CDC) highlights the severity of the crisis, reporting over 2.8 million antibiotic resistant infections annually, resulting in more than 35,000 deaths.[2] Traditional clinical surveillance for antibiotic resistant infections involves testing bacteria isolated from clinical specimens to assess antibiotic susceptibility. This method is resource-intensive and only reflects antibiotic resistance genes (ARGs) in individuals seeking medical treatment, limiting our understanding of the true burden and diversity of AMR in the broader community.

Various anthropogenic factors contribute to the levels of AMR circulating in communities, but the relative influence of these factors remains unclear. Improper use of antibiotics in both healthcare and agriculture has long been recognized as a major driver of antibiotic resistance, with overprescription and misuse accelerating the development of resistant bacteria.[3] However, recent studies have increasingly emphasized the role of socioeconomic factors in shaping resistance patterns.[4,5] Disparities in access to healthcare, sanitation, and education can influence not only antibiotic use but also the effectiveness of infection control measures. In a global analysis, Collignon et al. reported limited association between clinical isolates antimicrobial resistance levels and antibiotic consumption but found stronger association with socioeconomic factors such as GDP per capita, education, infrastructure, and public healthcare spending.[5] Similarly, a global analysis of sewage metagenomes by Gupta et al. and Lee et al. identified significant correlations between resistance levels and socioeconomic factors.[6,7] Additionally, global travel has also been found to play a key role in spreading resistant pathogens, facilitating the dissemination of resistance across borders.[8,9] These interconnected drivers highlight the need for a comprehensive examination of multiple determinants of antibiotic resistance.

Recent advances in wastewater monitoring for SARS-CoV-2 and other viral pathogens have established data collection infrastructure that can be leveraged to monitor various markers of infectious disease and human health, including ARGs, in a timely and comprehensive manner.[10] Wastewater has consistently been recognized as a reservoir for ARGs; however, most research to date has focused on the fate of ARGs across the wastewater treatment processes. While there are a limited number of studies analyzing ARGs in wastewater to gain insights for antibiotic resistance in communities, these studies either focus on global spread, characterizing and comparing the wastewater resistome globally,[4,6,11,12] or are limited to one city or state.[13–15] Additionally, previous wastewater ARG studies have mostly been conducted using shotgun metagenomics. While metagenomic approaches are useful in understanding ARG diversity, they cannot detect rare targets and do not provide accurate quantitative estimates. In contrast, digital PCR (dPCR) has the advantage of being highly sensitive and quantitative, allowing for normalized abundance estimates and comparison across samples in time and space.

In this study, we use dPCR to quantitatively measure the abundance of 11 clinically relevant ARGs that confer resistance to beta-lactams (*CMY, CTX-M, KPC, NDM, mecA, OXA, TEM, VIM*), colistin (*mcr-1*), tetracycline (*tetW*), and vancomycin (*vanA*) from 163 wastewater sites to generate a cross-sectional data set of the U.S. By linking publicly available secondary data, we investigate potential determinants of wastewater ARG concentrations, including antibiotic use, social vulnerability, the size of animal agriculture operations, and the locations of healthcare facilities and airports. Finally, using a subset of our data as a training set, we construct a random forest model to predict antibiotic resistance gene prevalence in all counties across the U.S.

Results

Clinically relevant ARGs are at high concentrations in wastewater across the U.S.
We measured the concentration of 11 antimicrobial resistance genes (ARGs) across 163 sites in the U.S, spanning 40 states for a cross-sectional data set (**Figure 1A**). An average of 2.7 samples were collected per wastewater treatment plant over a week and concentrations were averaged to obtain one representative concentration for each target per treatment plant (**Table S1**). We selected the targets based on the Centers for Disease Control Antibiotic Resistance Threats Report that identifies pathogen-resistance combinations as urgent, serious, and concerning based on their threat to human health.[1] Targets included genes that confer resistance to beta-lactams, colistin, tetracycline, and vancomycin. The ARGs *CMY, CTX-M, KPC, OXA-48, TEM,* and *tetW* consistently exceeded 10^{-7} copies per copy of 16S rRNA gene and were detected in all samples (**Figure 1B**). Among these, *tetW* exhibited the highest concentration across all samples, with minimal variation in its distribution. In contrast, *mecA, NDM, VIM, mcr-1,* and *vanA* displayed more variability in their distribution. Notably, *NDM* and *VIM* concentrations exhibited a broad distribution, ranging from undetected to $3x10^{-5}$ and $8x10^{-4}$ copies per 16S rRNA gene respectively. All targets in this group showed either non-detects (i.e. all wastewater samples collected during the week of sampling were non-detects for the WWTP) or partial detects (i.e., at least one sample collected for the WWTP during week of sampling was a non-detect). *Mcr-1* had the highest number of non-detects (38.3% of sites) and partial detects (32% of sites), followed by *vanA* (24.5% of sites non-detects and 22.7% of sites partial detects) and *NDM* (21.5% of sites non-detects and 16.6% of sites partial detects). Overall, most ARGs in wastewater were positively correlated with each other (Spearman's correlation coefficients ranged from 0.14 to 0.87, all $p < 0.05$), with beta-lactamase genes showing some of the strongest pairwise correlations, except for *OXA-48* and *mecA* (**Figure S1A**).

To summarize the resistance burden, we calculated z-scores based on the distributions of each gene normalized by 16S rRNA gene concentrations, generating burden scores for each resistance class and total resistance. Although more than the presence of ARGs is needed to estimate burden, we will use this term here to mean antibiotic resistance gene prevalence in the community's wastewater. Burden z-scores were highly correlated (**Figure S1B**). Comparison of these burden scores across census regions revealed significant regional differences in both total and beta-lactamase resistance burden (Kruskal-Wallis test, $p < 0.05$; **Figure 1C**). Specifically, the Northeast and the South had significantly higher total resistance burden

compared to the Midwest and the West. A similar trend was observed for beta-lactamase resistance. In contrast, the West and the South demonstrated significantly higher colistin resistance burdens than the Midwest and the Northeast (**Figure 1D**; Post-hoc Conover-Iman test, $p < 0.05$, Benjamini-Hochberg correction applied). For individual beta-lactamase genes, significant regional differences were observed without a clear trend (**Figure 1D**). For instance, the Midwest had significantly higher *CTX-M* levels but lower *KPC* concentrations. Regional differences among all ARGs can be seen in **Figure S2.** The two genes with the highest national variability, *NDM* and *VIM*, displayed distinct patterns: *NDM* showed significant regional differences, with the South and the West having higher concentrations than the Midwest and the Northeast, while *VIM* exhibited similar variability across all regions.

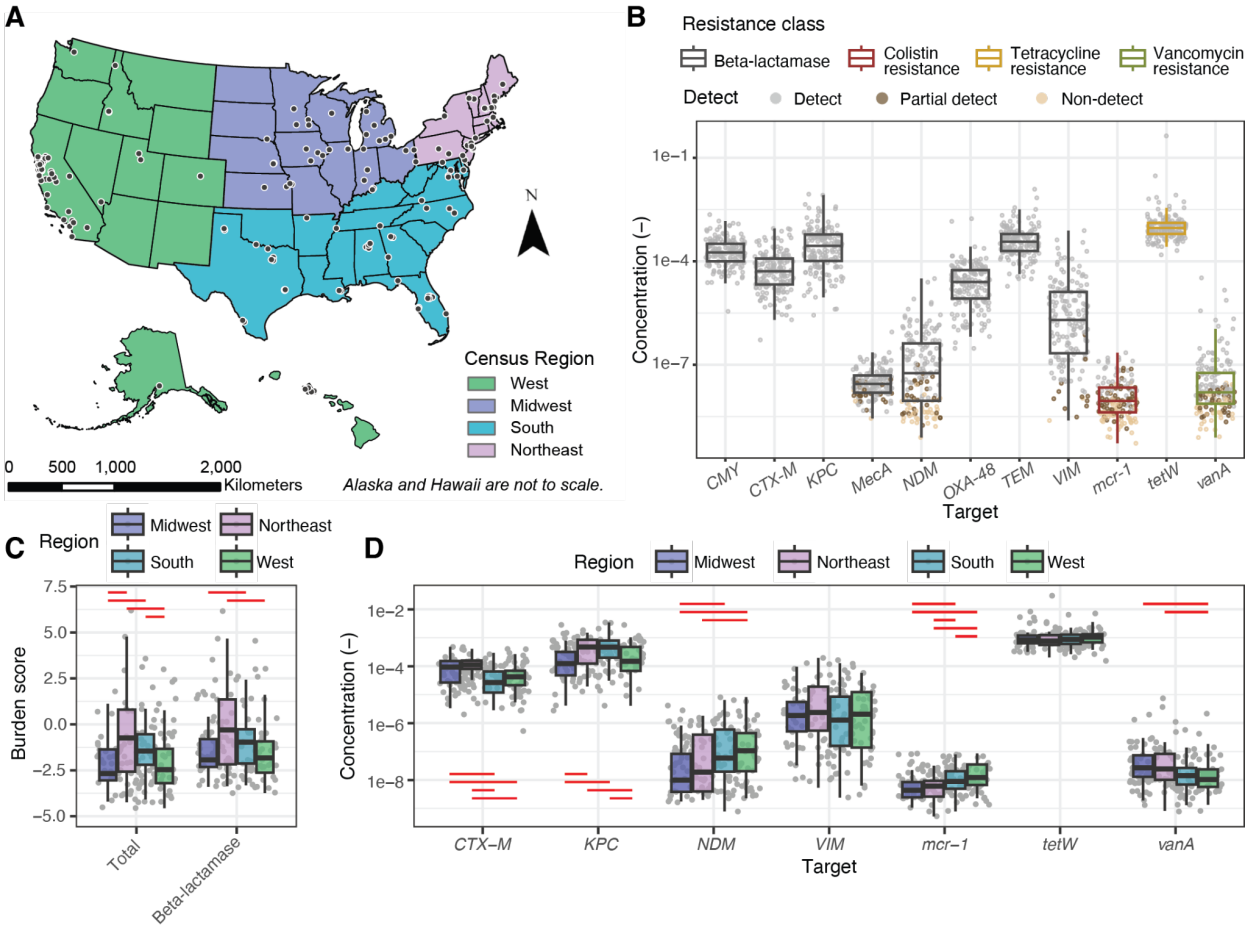

Figure 1. Overall and regional distribution of ARGs and antibiotic resistance burden. A) Sample locations (black circles) for this study divided into census regions. **B)** ARG concentration normalized by 16S rRNA gene for all sample sites. **C)** Total resistance burden and beta-lactamase resistance burden calculated as z-scores for each census region. **D)** Select beta-lactamase genes and colistin, tetracycline, and vancomycin resistance genes in concentrations normalized by 16S rRNA gene across census regions. Each symbol in panels B, C, and D represents a wastewater treatment plant. If the gene was undetected, half of the theoretical measurement limit was substituted as measured value. The median is shown by the line inside the box with the 25th and 75th percentile represented by the lower and upper

boundary of the box. Bottom and top whiskers show 1.5 x interquartile range. Red lines indicate a significant pairwise difference between regions measured by the Conover-Iman post-hoc test with Benjamini-Hochberg correction applied ($p < 0.05$).

Determinants of wastewater antibiotic resistance burden

Using publicly available datasets, we investigated the influence of specific anthropogenic or socioeconomic factors on the antibiotic resistance burden observed in wastewater. Values for secondary variables for sites were determined at either the sewershed scale or for the county predominantly intersecting a sewershed (see Supporting Information). Several notable associations were identified. Colistin resistance measured in wastewater was associated with overcrowding, speaking limited English, and certain race and ethnicities ($p < 0.05$ after Benjamini-Hochberg adjustment; **Figure 2A**). Overall beta-lactamase resistance was associated weakly with many variables but specific beta-lactamase genes (*CMY*, *KPC*, *NDM, TEM)* showed stronger correlations with socioeconomic factors such as being uninsured, being under housing burden, overcrowding, and lack of a high school diploma, as well as speaking limited English and race and ethnicity variables ($p < 0.05$ after Benjamini-Hochberg adjustment; **Figure 2B**). Among these four genes*, NDM* exhibited the greatest number of significant correlations. Antibiotic use from Epic Cosmos and the number of animals had little to no correlation with the resistance burden. These correlations were adjusted with Benjamini-Hochberg method to control the false discovery rate (**Figure 2**); unadjusted correlation coefficients are provided in **Figure S3**.

Total antibiotic resistance burden score differed significantly based on the presence of hospitals and whether the number of nursing homes in the sewershed exceeded the median (7) (**Figure 2C**). While colistin resistance gene concentration did not differ based on the presence of hospitals, higher *mcr-1* concentrations were associated with the presence of airports, higher number of nursing homes, higher population density, and urbanicity (**Figure 2D**). When examining individual beta-lactamase genes, some genes, such as *CMY* showed no significant difference with presence of points of interest. In contrast, other genes like *NDM*, *KPC*, *TEM*, and *VIM* showed significant differences based on presence of hospitals and the number of nursing homes (**Figure 2E, F**; **Figure S4E, H, I**). Notably, higher *NDM* concentration was the only beta-lactamase gene that was associated with the presence of airports, higher population density, and urbanicity (**Figure 2F**).

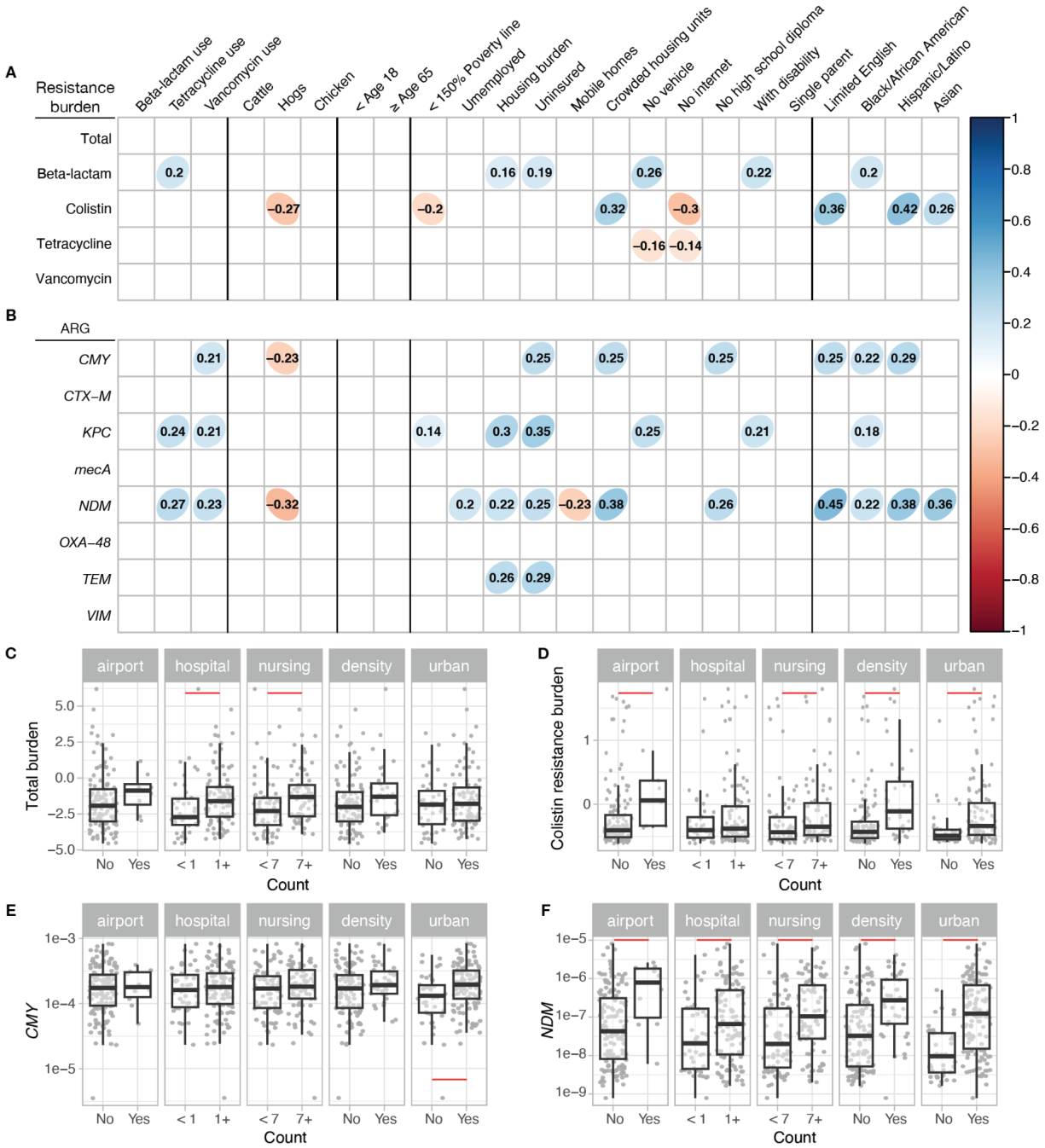

Figure 2. **Association between antibiotic resistance found in wastewater and potential determinants.** Spearman's correlation coefficient among potential determinants of resistance burden and **A)** antibiotic resistance burden score and **B)** beta-lactamase gene concentrations measured in wastewater and normalized by 16S rRNA gene ($p < 0.05$ with Benjamin-Hochberg correction applied). Bivariate analysis of points of interest, population density, and urbancity for **C)** total antibiotic resistance burden and **D)** colistin resistance burden measured as z-scores and **E)** *CMY* and **F)** *NDM* concentration measured in wastewater normalized by 16S rRNA. For **C-F**, the median is shown by the line inside the box with the 25th and 75th percentile represented by the lower and upper boundary of the box. Bottom and top whiskers show 1.5 x

interquartile range. Red lines indicate a significant pairwise difference between regions measured by the Wilcoxon post-hoc test with Benjamini-Hochberg correction applied ($p < 0.05$). Specific statistics of Wilcoxon post-hoc test shown in **Table S7**.

Non-clinical indicators predict antimicrobial resistance concentrations

Using publicly available datasets of non-clinical determinants of antibiotic resistance, we developed random forest models to predict wastewater ARG concentrations across all U.S. counties. The random forest models presented here had one tuning parameter: the number of variables randomly sampled at each split, which ranged from 2 to 25. For all ARGs, the model that minimized RMSE considered 2 variables at each split, so all subsequent results are presented with this configuration. Overall, the predictive performance for individual genes, indicated by R^2, ranged from 0.22 to 0.41 (**Figure S5**). The *mcr-1* model, showing colistin resistance, achieved one of the highest predictive performance with an R^2 of 0.40, indicating it could explain 40% of the variance in colistin resistance across counties based solely on the secondary dataset. The models for *KPC* and *NDM*, both conferring carbapenem resistance and classified as urgent threats by the CDC, showed similar predictive performance with R^2 values of 0.30 and 0.31 respectively. The vanA model, showing resistance to vancomycin, had an R^2 of 0.36 (**Figure 3**). In each case, we can explain a significant amount of the underlying variation in our concentration measurements based solely on non-clinical indicators. Variables identified to be the most important across all individual ARG prediction models were proportion of the population speaking limited English, being under housing burden, being uninsured, certain race and ethnicities, and urbanicity of the sewershed (**Figure 4**). Most ARGs had many different variables contributing similarly to the prediction model, except *mcr-1*, which had three prominent variables: proportion of people who live in a crowded housing unit, speak limited English, and do not have a high school diploma.

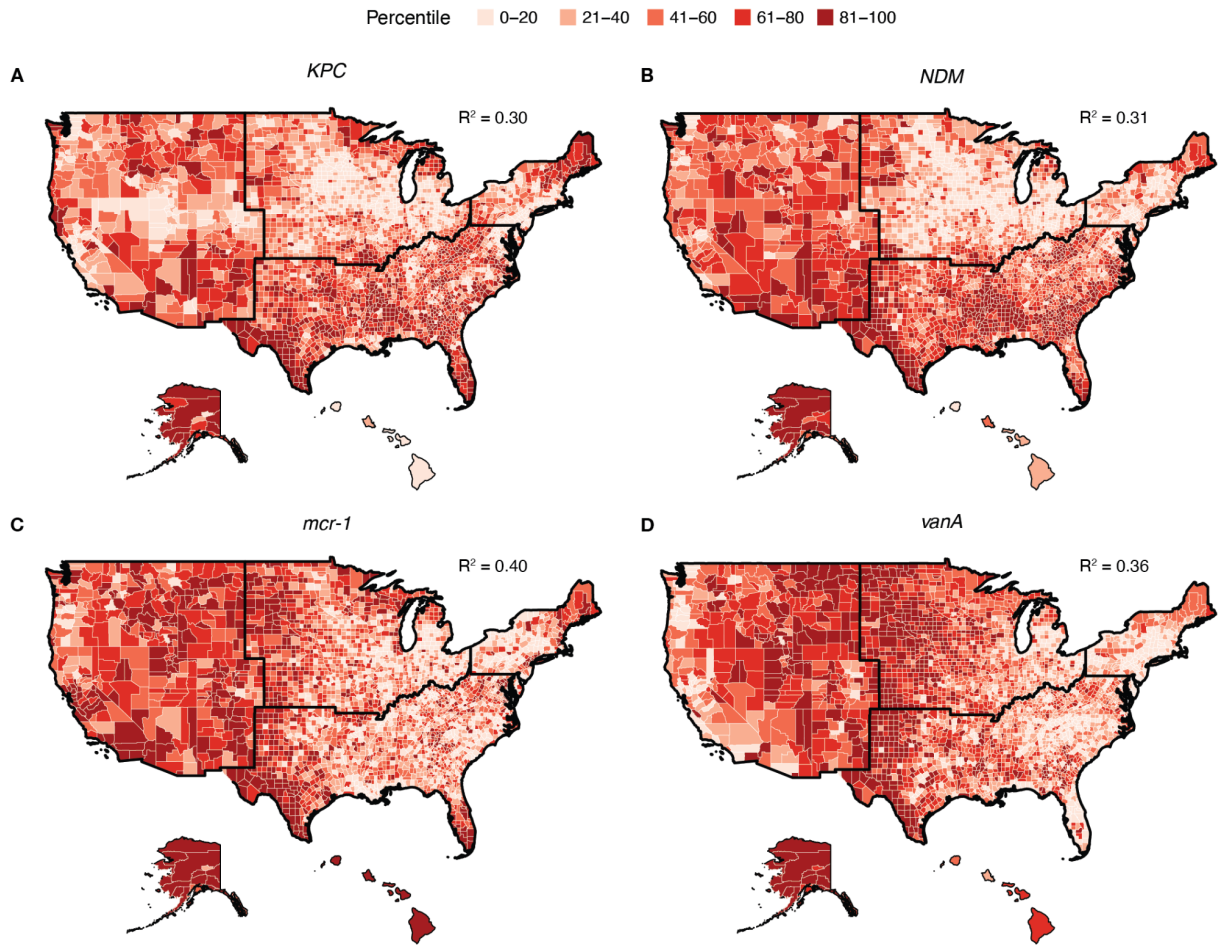

Figure 3. Random forest modeling of antibiotic resistance gene concentrations across the United States using secondary data. Predicted **A)** *KPC* (carbapenem resistance; R^2: 0.30), **B)** *NDM* (carbapenem resistance; R^2: 0.31), **C)** *mcr-1* (colistin resistance; R^2: 0.40), and **D)** *vanA* (vancomycin resistance; R^2: 0.36) concentrations across the United States. The predicted concentration, normalized by 16S rRNA, is visualized as percentiles, with darker colors indicating higher concentrations. Random forest models are trained using the secondary data in **Table 1**, with the exception of antibiotic prescription rates. Alaska and Hawaii are not to scale.

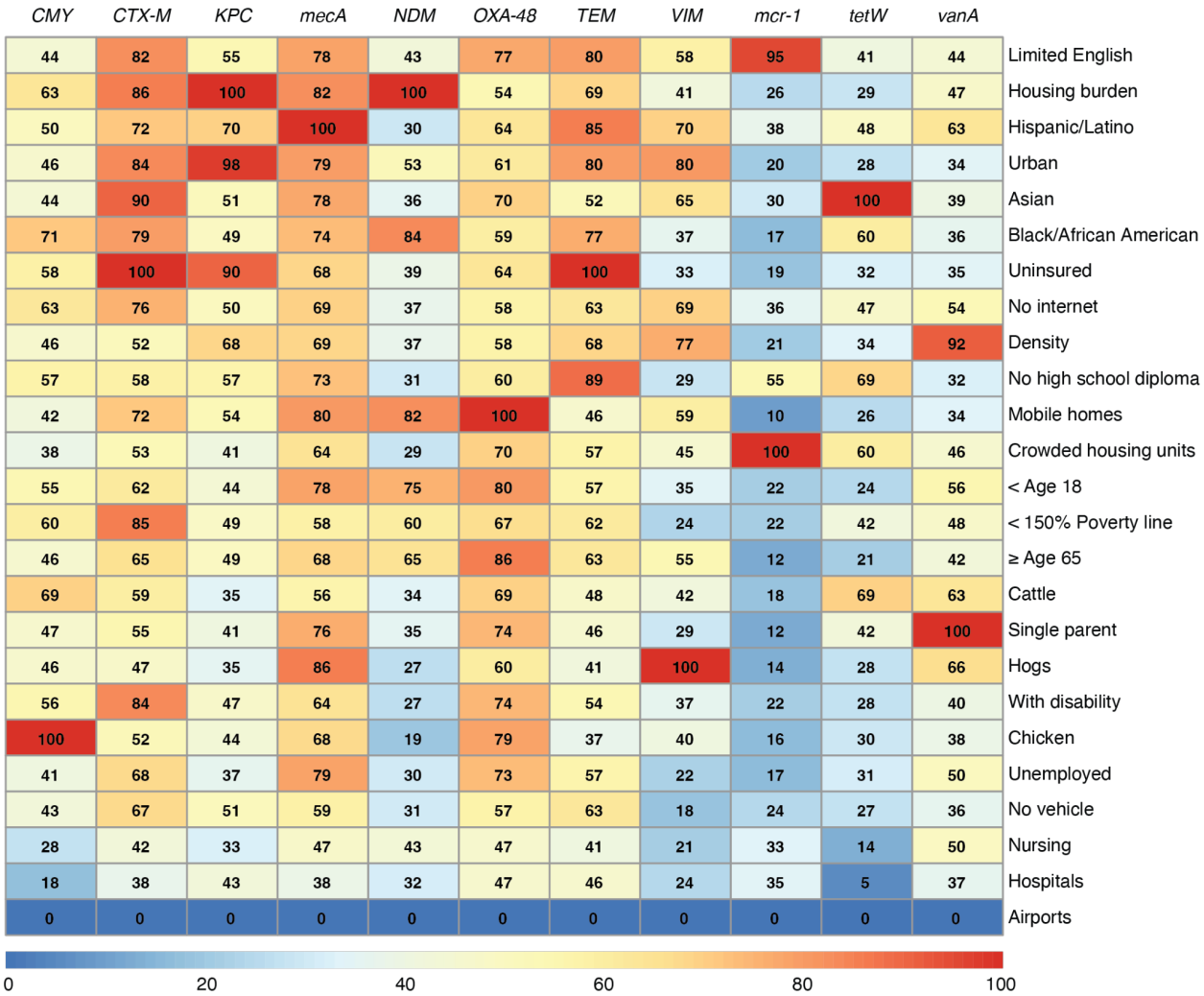

Figure 4. Summary of variable importance for random forest prediction models for each ARG. The rows represent different variables used in the random forest models, sorted by their overall importance, summed across all genes, from highest to lowest. The color gradient reflects the magnitude of the relative importance of each variable for each ARG prediction model, with red indicating higher importance. Individual prediction model importance plots are shown in **Figure S6**.

Discussion

This study establishes a quantitative baseline for the current concentrations of ARGs in wastewater across the U.S., and provides new insights for understanding the complex factors driving resistance gene abundance circulating at the population level. We present evidence that the antibiotic resistance burden disproportionately affects low-income and disadvantaged communities. Increased levels of ARGs in wastewater were associated with limited access to health insurance, housing burden, and overcrowding. Health insurance access and housing burden were also among the most frequently identified socioeconomic variables of importance

in our random forest prediction model, along with lack of a high school diploma and living in a mobile home. In particular, colistin resistance and certain beta-lactamase genes (*CMY*, *KPC, NDM*, and *TEM*) exhibited some of the strongest positive associations with these socioeconomic factors. These findings align with previous studies linking socioeconomic status to risk of AMR, based on clinical isolates or sewage metagenomes.[6,7,16–18] For example, Cooper et al. found significant spatial correlation in Dallas-Fort Worth, Texas, metropolitan area between AMR organisms (particularly AmpC beta-lactamase and methicillin-resistant *Staphylococcus aureus*), and high levels of the Area Deprivation Index, which includes factors like income, education, employment, and housing quality.[16] In our study, *CMY*, an AmpC beta-lactamase gene, was associated with being uninsured, overcrowding, and low education levels, and low income status was one of the most important predictor variables in the random forest model. Race was also significantly and positively associated with many of the resistance genes, and minority race variables were the most frequently identified important variable across all potential determinants in the random forest model. Race can be correlated with systemic disadvantages, such as limited access to healthcare and education, which are well-documented risk factors for resistant infections.[17,19] For instance, Black communities and areas with high rates of poverty often experience higher rates of MRSA infections.[17,18] Overall, these findings underscore that the risk of antibiotic resistance infections is linked to socioeconomic disparities.

We also identified international travel and recent immigration as significant drivers of increased ARG burden in domestic wastewater. *NDM* and colistin resistance (*mcr-1*) were significantly associated with the number of airports and the proportion of the population who speak limited English. Previous research has shown that international travelers frequently acquire AMR organisms, contributing to the global spread of resistance genes. Depending on travel destination and behavior, the acquisition rate of extended-spectrum beta-lactamase-producing *Enterobacterales* has been shown to be as high as 75%.[20] A recent preprint by Sridhar et al. also found *mcr*-mediated colistin-resistant *Enterobacterales* in metagenomes from post-travel stool samples, with the highest associations found among travelers to South America and South-Eastern Asia.[21] Here, *mcr-1* also showed significant associations with the presence of Hispanic/Latino and Asian populations. In the U.S., high cost of healthcare and lack of insurance in low-income populations can be a crucial reason for international travel. For instance, a survey of adults in Los Angeles by Macias and Morales found that many low-income and uninsured individuals seek low-cost health care and medications in Mexico to meet their urgent health care needs, despite the significant burden of travel and cost.[22] This also ties international travel to similar socioeconomic factors identified in our study that correlate with wastewater ARG concentrations. In addition, non-prescription use of antibiotics may be influenced by cultural differences among recent immigrants. In a cross-sectional survey conducted by Corbett et al. in Colorado, significant gaps were found in antibiotic resistance awareness between English and non-English speakers, with ethnicity serving as a predictor of knowledge, attitudes, and awareness regarding antibiotic use and resistance.[23] Similarly, Hawkes et al. identified race to be a major factor related to distrust in the healthcare system, leading to non-prescription antibiotic use.[24] However, there is some evidence suggesting that Americans of international heritage may have different attitudes and behaviors towards antibiotics use than more recent immigrants.[25] Therefore, the correlation between race and ARG

prevalence in wastewater may be driven by a combination of socioeconomic factors and immigration status, emphasizing that societal and cultural patterns influence antibiotic use.

Antibiotic use itself showed weak or no correlation with presence of ARGs. This lack of correlation may be due to discrepancies between the amount of prescribed antibiotics in healthcare settings and actual antibiotic consumption across humans and animals. Riquelme et al. found that global trends in sewage antibiotic concentrations, measured with LC-MS/MS, did not align with trends predicted from reported human antibiotic sales data.[26] While this discrepancy may be less pronounced within the U.S., it underscores a disconnect between antibiotic sales data and actual consumption, which likely reflects large antibiotic usage in food and companion animals that is also expected to drive antibiotic resistance.[27] Another potential factor contributing to this disconnect could be non-prescription antibiotic use, which is not captured by prescription data. Although the overall outpatient antibiotics use in the U.S. has gradually declined (potentially indicating more careful prescribing from clinicians),[28] greater attention must be directed toward antibiotic use in animals and non-prescription usage, both of which cannot be addressed through hospital-based antibiotic stewardship programs. Depending on the population characteristics, non-prescription antibiotic use has been shown to vary from 1% to 66%.[29] Importantly, our ability to predict wastewater ARG concentrations using primarily non-clinical indicators reinforces our finding that antibiotic resistance is driven by broader social, environmental, and infrastructural factors beyond just prescribed antibiotic use. These findings suggest that current antibiotic stewardship efforts may be insufficient in addressing the broader societal dynamics that fuel the global antibiotic resistance crisis.

By examining the quantitative levels of individual ARGs across the nation, we observed varying degrees of national variation, which informs the utility of certain genes as indicators of broader resistance patterns. For instance, genes like *tetW*, which exhibit minimal variation nationwide, may not offer much insight into the resistance burden since its distribution is relatively consistent. On the other hand, genes with larger variation, such as *NDM* and *VIM*, may provide more meaningful information about resistance dynamics. Notably, across individual genes we observed significant differences in regional variation and correlation with potential determinants. The maps shown in **Figure 3** reinforce regional differences in antibiotic resistance gene concentrations, and the variable importance ranking (**Figure 4**) gives potential insight for these differences. The use of z-scores to aggregate resistance burden allowed for a standardized comparison, but may obscure gene-specific insights that were obtained when examining individual genes, consistent with other summary metrics used in previous studies.[30,31] This highlights the value of measuring individual genes and modeling their abundance separately, and underscores the difficulty of developing an overall resistance burden indicator. The model also performed less effectively with z-scores representing overall resistance (**Figure S7**). This aligned with the correlation analysis where more insights were gained by examining individual genes, underscoring a need for granular analysis to understand community-level resistance dynamics.

This study has several limitations. Regional coverage of wastewater samples received was uneven, with some areas, such as California, overrepresented. However, the study covers 40

out of 50 states of the U.S and includes 163 sites. Additionally, matching sewershed boundaries to counties may have introduced inaccuracies when linking resistance data to local characteristics. It is also important to note that this is a cross-sectional analysis. In a global analysis of sewage metagenomes, Munk et al. found no obvious trends in time and only a modest effect of seasonality;[30] therefore we do not anticipate a substantial variation in these outcomes across seasons. However, if any factors lead to a delayed response in the wastewater signal, as recent studies suggest could occur with antibiotic use,[32,33] this study would not be able to assess such effects. While our use of ddPCR does not reveal the genomic context of the ARGs detected, all the targets in this study are either plasmid-mediated resistance genes or can be carried on mobile genetic elements, which increases their risk of transmission into human microbiota. Our analysis is generalizable to the U.S. and potentially other high-income settings; other studies in low and middle income countries would likely identify different risk factors.[7] Lastly, the R^2 values of the random forest model suggest that the chosen set of determinants explains a substantial portion of the variation in ARG concentrations, but there are other unmeasured factors influencing resistance patterns across the country that were not captured in this study.

This study presents a quantitative analysis of clinically relevant ARGs in wastewater, offering a valuable tool and baseline for understanding and monitoring resistance dynamics across the U.S. The use of wastewater to study ARGs allows for real-time, population-level insights into the presence and distribution of antibiotic resistance, supplementing clinical data that may be underestimating the resistance burden. The robust baseline established by this study can be used to track changes in resistance gene prevalence over time, which could help identify emerging resistance trends, track the impact of interventions, and better predict the trajectory of resistance in the population. One of the significant merits of this study is its sensitivity in detecting rare but clinically important resistance genes enabled by the use of digital PCR. This method enables the identification of low-abundance resistance genes that can be missed in studies relying on non-targeted sequencing methods like metagenomics.[34,35] This enhanced sensitivity is critical for accurate detection of emerging resistance genes, such as colistin resistance or certain beta-lactamase genes that are low abundance, which can have profound implications for clinical treatment strategies. Our findings provide important insights into the complex relationship between socioeconomic factors and antibiotic resistance, suggesting that antibiotic stewardship is not sufficient on its own but improving access to healthcare and raising the standard of living in the U.S. is necessary to reduce antibiotic resistance.

Materials and Methods

ddPCR assay optimization

We chose probe-based qPCR assays in literature that have been utilized for detection of the target ARGs in wastewater to test on the ddPCR platform (**Table S2**). The assays were initially tested on gblocks (IDT; sequences provided in the SI) of the target region of each ARG as either a singleplex (FAM) or a duplex (HEX) assay with an annealing temperature gradient ranging from 55°C to 65°C in order to identify an annealing temperature that yielded sufficient

separation between negative and positive droplets (**Figure S8**). Subsequently, the assays were tested using templates at four different dilutions (neat, diluted in molecular grade water at 1:10, 1:100, 1:1000) of wastewater samples collected from seven wastewater treatment plants in the U.S. with distinct geographical characteristics to identify the dilution factor suitable for each target ARG. Consequently, we categorized the target ARGs into two groups based on their expected concentrations in wastewater (high versus low) in order to multiplex the assays into two 6-plex assay mixes that utilized an annealing temperature of 58°C.

Sample collection

Wastewater treatment plant (WWTP) staff provided either "grab" samples from the primary clarifier or 24-hour composite samples from the headworks (**Table S3** shows which WWTP collected each type). The "grab" samples in this case are solids collected in the primary clarifier over 1-8 hours, representing composite community wastewater solids. The samples were then stored at 4°C, shipped to the laboratory overnight, and processed within 48 hours of receipt at the laboratory. A total of 443 samples were collected and analyzed for this study (**Figure 1A, Table S1**).

Pre-analytical processing and nucleic-acid extraction

Methods for pre-analytical processing and nucleic-acid extraction and purification have been described in detail by Boehm et al., so are briefly summarized here.[36] Wastewater solids were obtained from the sample by use of Imhof cones and centrifugation and then suspended in a buffer (DNA/RNA shield, Zymo Research, R1100-250, California, USA) containing bovine coronavirus (as a positive extraction control; Calf-Guard Cattle Vaccine, Ohio, USA) at a concentration known to reduce the potential for inhibition in the analytical steps.[37] The samples were homogenized after addition of grinding balls (5/32-in., OPS Diagnostics, New Jersey, USA) and then nucleic-acids extracted and purified from 6-10 aliquots of the supernatant to obtain 6-10 300 µl replicate extracts from each sample (see **Table S3** for whether 6 or 10 replicates used for the plants). Nucleic-acid extraction and inhibitor removal was conducted using Chemagic Viral DNA/RNA 300 kit H96 (Revvity, CMG-1033-S, Massachusetts, USA) for the Chemagic 360 and Zymo OneStep-96 PCR Inhibitor Removal kit (Zymo Research, D6030, California, USA). The methods for nucleic acid extraction are described step-by-step in protocols.io.[38] The first two aliquots of every sample were chosen to be analyzed for ARGs. Extraction-negative controls (water) were extracted using the same protocol as the homogenized samples. Nucleic-acids were used immediately for quantification of ARGs. A portion of the solids were dried to determine their dry weight.[36]

Droplet Generation and Quantification

The probe used for each assay was labeled with one of six fluorophores (FAM, HEX, Cy5, Cy5.5, ROX, and/or ATTO590) and run in a 6-plex assay using the probe-mixing approach (**Table S2**). All targets were multiplexed and run together based on being High concentration (samples diluted 1:100 with ultrapure H_2O) or Low concentration (undiluted samples). Two wells were run per reaction using the first two nucleic-acid extractions. Digital droplet PCR was performed on 20 µL samples from a 22 µL reaction volume, which consisted of a 5.5 µL template. Final primer concentrations were 900 nM and probe concentrations were 250 nM for

all ARG targets. All ARG master mixes were prepared using ddPCR Supermix for Probes (BioRad #1863010, California, USA). Droplet generation was performed using an Automated Droplet Generator (BioRad #1864101, California, USA). Quantification was performed using a QX600 Droplet Digital PCR system (BioRad #17007769, California, USA). All assays were performed at an annealing temperature of 58°C. Thermocycling conditions were as follows: 5 minutes at 95°C, followed by 40 cycles of 30 seconds at 95°C and 1 minute at 58°C, followed by 10 minutes at 98°C, and an infinite hold at 4°C. Each plate run included no template controls, extraction negative controls, extraction positive controls, and positive controls, which all performed as expected. Extraction negative controls used ultra-pure water, extraction positive controls used spiked BCoV as previously described, and positive controls were synthetic IDT gblocks which included the target amplicon. No template controls and extraction negative controls were negative for all targets in all runs. Extraction positive controls and template positive controls were positive in all runs. For each well in ddPCR, a minimum of 10,000 droplets was required and at least three positive droplets were required to be quantifiable. If a target was undetected from a diluted sample, the measurement was redone or excluded (one *mecA*, six *OXA-48*, and one *mcr-1* measurement excluded). In cases of duplicate measurements at different dilutions, the value with the smaller dPCR confidence interval was selected.Two replicate wells were merged per measurement. Thresholding for each target was performed automatically using the QX Manager Software Version 2.2 (BioRad, #168032, California, USA).

After quantification of ARGs, the leftover nucleic acid replicates were combined and stored at -80°C for 17 days to 21 days prior to being used as templates for measuring high copy number bacterial target, 16S rRNA. 16S rRNA genes were quantified using ddPCR EvaGreen Supermix (BioRad, #1864034, California, USA) and QX200 Droplet Digital PCR system (BioRad, #1864001, California, USA). EvaGreen is an intercalating dye that binds to double stranded DNA. Final primer concentrations were 100 nM for 22 µL reaction volume, 5 µL of which was the template. 16S rRNA quantification was run in singleplex reactions using template diluted 1:50,000 with UltraPure H_2O. One well was run per reaction using the combined template from the quantification of ARG genes. Thermocycling conditions were as follows: 5 minutes at 95°C, followed by 40 cycles of 30 seconds at 95°C and 1 minute at 55°C, followed by 5 minutes at 4°C, 5 minutes at 90°C and infinite hold at 4°C. Each plate run included no template controls (NTCs). The NTCs resulted in a small amount of positive droplets (average 16 positive droplets) and therefore any sample concentration on the same plate with less than an order of magnitude difference was excluded from analysis (three 16S rRNA measurements excluded). All wells had more than 10,000 partitions. Thresholding for each target was performed automatically using QX Manager Software Version 2.2.

Additional details associated with the Minimal Information MIQE Experiments (dMIQE) for digital droplet PCR reporting are described here.[39] The average (standard deviation) number of droplets for the quantification of ARGs was 34225 (7143; 2 merged wells) and for bacterial count was 17997 (1443; 1 well). The volume of partitions, as reported by BioRad, was 0.000795 µL. The mean and standard deviation of copies per partitions for each target is shown in **Table S6**. Median bovine coronavirus (BCoV) recoveries (BCoV measured in sample / BCoV

measured in DNA/RNA shield without sample) were, on average 1.06 (25th percentile = 0.71, 75th percentile = 1.3) across all samples. Recovery higher than 100% is likely due to an underestimation of the amount of BCoV spiked into the samples. Based on the recovery, we concluded that there was no gross inhibition of the samples. After QA/QC, we had 443 different samples and 5309 measurements that represented 163 WWTPs in 40 states and 134 cities in the U.S.

Demographic, socioeconomic, and environmental data

Antibiotic resistance may be influenced by various demographic, socioeconomic, and environmental determinants—in addition to antibiotic usage. Demographic and socioeconomic factors affect a community's capability to respond and adapt to external hazards and stressors, such as public health emergencies. A community's status as urban or rural or its population density may affect its access to and quality of healthcare services. Proximity to major points of entry (e.g., airports) and health care facilities (e.g., hospitals, nursing homes) may facilitate the importation or transmission of public health threats in communities. Antibiotics are also commonly used for agricultural purposes, and ARGs may be transmitted zoonotically within communities. Therefore, we characterized sewersheds based on several demographic, socioeconomic, environmental, and antibiotic usage determinants (**Table 1**) to assess their relationship with wastewater concentrations of ARGs. Most WWTPs (n = 94 of 163) provided a sewershed boundary. For WWTPs that did not provide a sewershed boundary (n = 69 of 163), we approximated a boundary based on the zip code(s) serviced by the WWTP and 2023 USA ZIP Code Boundaries published by Esri (Source: TomTom, US Postal Service, Esri).[40] Because sewershed boundaries do not align with traditional boundaries (e.g., counties or census tracts), we approximated sewershed-level estimates for each determinant in **Table 1** as described in the Supporting Information. The demographic and socioeconomic determinants we selected are based on the 2022 Social Vulnerability Index (SVI) from the US CDC as social vulnerability describes the degree to which demographic and socioeconomic factors affect a community's resilience.[41] Antibiotic usage was determined using Epic Cosmos (Epic Cosmos, Epic Systems Corporation, Wisconsin).[42]

Table 1. Demographic, socioeconomic, environmental, and antibiotic usage variables that may influence antibiotic resistance

Sewershed Characteristic	Description
< 150% Poverty line[a]	Proportion of the population for whom poverty status is determined below the 150% poverty line
Unemployed[a]	Proportion of the civilian labor force age 16+ years that is unemployed
Housing burden[a]	Proportion of occupied housing units that are housing cost-burdened with an annual income <$75K
No high school diploma[a]	Proportion the population age 25+ years with no high school diploma

Uninsured[a]	Proportion of the civilian noninstitutionalized population that is uninsured
≥ Age 65[a]	Proportion of the population age 65+ years
< Age 18[a]	Proportion of the population age 17 years and younger
With disability[a]	Proportion of the civilian noninstitutionalized population with a disability
Single parent[a]	Proportion of single-parent households with children <18 years
Limited English[a]	Proportion of persons age 5+ years who speak English "less than well"
Black/African American[a]	Proportion of the population that is Black/African American, not Hispanic or Latino
Hispanic/Latino[a]	Proportion of the population this is Hispanic or Latino
Asian[a]	Proportion of the population that is Asian, not Hispanic or Latino
Mobile homes[a]	Proportion of housing units that are mobile homes
Crowded housing units[a]	Proportion of occupied housing units with more people than rooms
No vehicle[a]	Proportion of occupied housing units with no vehicle
No internet[a]	Proportion of households with no internet
Urban[b]	Proportion of the area that is classified as urban
Density[a]	Population density (people per square kilometer)[g]
Airport[c]	Number of major airports
Hospital[d]	Number of hospitals
Nursing[d]	Number of nursing homes
Cattle[e]	Number of cattle[g]
Chicken[e]	Number of chickens[g]
Hogs[e]	Number of hogs[g]

Beta-lactam use[f]	Proportion of encounters with beta-lactam antibiotics (penicillins, cephalosporins, and carbapenems) dispensed[g]
Tetracycline use[f]	Proportion of encounters with tetracycline antibiotics dispensed[g]
Vancomycin use[f]	Proportion of encounters with vancomycin antibiotics dispsensed[g]

Refer to the Supporting Information for further information about how values for each characteristic were determined for the sewersheds included in this study.
[a] Data source: 2022 5-year American Community Survey[43]
[b] Data source: 2020 US Census Bureau urban-rural designations[44]
[c] Data source: Esri (source: Federal Aviation Administration's National Airspace System Resource Aeronautical Data Product)[45]
[d] Data source: US Department of Homeland Security's Homeland Infrastructure Foundation-Level Data database[46]
[e] Data source: 2022 US Department of Agriculture National Agricultural Statistics Service[47]
[f] Data source: Epic Cosmos (Encounters between June 1, 2023–June 1, 2024)[42]
[g] The value for these characteristics represents the predominant county of the sewershed (i.e., the county that the majority of sewershed is located in). The value for all other characteristics represents the sewershed area.

Data analysis

Analysis was conducted using R (Version 4.3.0). Outlier detection was performed using Rosner's test on 16S rRNA to exclude significant outliers. One 16S measurement was excluded. Each time point was treated as a replicate for each wastewater treatment plant and therefore averaged to obtain a final average concentration of each target. Samples were categorized as non-detect (ND) if none of the replicates detected the target, detect (D) if all replicates detected it, and partial detect (PD) if at least one of the replicates did not detect the target. For non-detect values, half of the lower measurement limit was calculated and assigned as a substitute value. Z-scores were calculated for each target by comparing the concentration to the overall distribution across all samples, using the mean and standard deviation. The AMR burden was calculated by summing up the z-scores of all ARG targets. The beta-lactamase burden was computed by summing up the z-scores of *CMY*, *CTX-M*, *KPC*, *mecA*, *NDM*, *OXA-48*, *TEM*, and *VIM*. Data points with z-scores greater than 3 or less than -3 were excluded from analysis involving secondary datasets (eight *CMY*, seven *CTX-M*, five *KPC*, one *mecA*, two *NDM*, two *OXA-48*, three *TEM*, four *VIM,* eight *mcr-1*, one *tetW*, and two *vanA* measurements excluded).

For statistical analysis, the Shapiro-Wilk test was used to assess normality, and Spearman's correlation coefficient was employed to evaluate relationships between variables, as the data was not normally distributed. For determinants of resistance with insufficient distribution (number of airports, number of hospitals, number of nursing homes, population density, and urbanicity), the data was split into two groups to conduct a bivariate analysis using the Wilcoxon rank sum test to determine if there was a significant difference in wastewater ARG concentrations between the lower and upper values of potential determinants. Number of airports and hospitals were split into groups based on the presence of the facilities, and the number of nursing homes were split by the median number of nursing homes (median: 7). The

sewershed were considered urban if there was 50% or more area considered to be urban. Population density was split by 617 per km^2, which is the average number of people in cities in the U.S. The Kruskal-Wallis test was applied to assess differences across Census regions, followed by the Conover-Iman test for post-hoc analysis to identify significant pairwise differences. The Benjamini-Hochberg correction was applied when potential determinants were involved and when investigating regional differences to control for the false discovery rate.

Predictive modeling

Based on our secondary data's characteristics, we chose random forest modeling to predict wastewater ARG concentrations in every county in the United States. Random forest modeling allows us to use data with varied distributions and nonlinear relationships, is resistant to outliers, and provides interpretable feature importance.[48,49] We first aggregated our mean wastewater ARG measurements at the county level. If a county was represented more than once in the dataset (i.e. contained 2 treatment plants), one of the sets of measurements from that county was randomly selected to represent the entire county. This left 108 unique observations. All variables from Table 1, excluding antibiotic usage variables, were considered and obtained for each county in the U.S. The variables for number of airports, number of nursing homes, and number of hospitals were turned into a percentile rank before analysis. To train the models, we used the caret package (version 7.0-1) in R.[50] Ten-fold cross-validation was used to evaluate the models, and we varied the number of features at each split between 2 and 26. Models were selected based on minimizing root mean square error. Variable importance was assessed using the varImp() function from the caret package, which randomly permutes variables and assesses the impact on prediction accuracy (**Figure S6**).

Data availability

All data used for analysis are available on the Stanford Data Repository: https://purl.stanford.edu/vb318cm9509.

The antibiotic prescription data from Epic Cosmos could not be made publicly available and therefore are excluded.

Code availability

All analysis R code used for this manuscript are available on github: https://github.com/sooyeolkim/ww_arg

References

1. U.S. Centers for Disease Control and Prevention. *Antimicrobial Resistance Threats in the United States, 2021-2022*. https://www.cdc.gov/antimicrobial-resistance/data-research/threats/update-2022.html (2024).
2. Centers for Disease Control and Prevention (U.S.). *Antibiotic Resistance Threats in the United States, 2019*. https://stacks.cdc.gov/view/cdc/82532 (2019) doi:10.15620/cdc:82532.

3. Holmes, A. H. *et al.* Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet* **387**, 176–187 (2016).
4. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).
5. Collignon, P., Beggs, J. J., Walsh, T. R., Gandra, S. & Laxminarayan, R. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. *Lancet Planet. Health* **2**, E398–E405 (2018).
6. Gupta, S., Wu, X., Pruden, A., Zhang, L. & Vikesland, P. Global scale exploration of human faecal and sewage resistomes as a function of socio-economic status. *Nat. Water* **2**, 975–987 (2024).
7. Lee, J. *et al.* Longitudinal monitoring of sewershed resistomes in socioeconomically diverse urban neighborhoods. *Commun. Med.* **5**, 7 (2025).
8. Sridhar, S., Turbett, S. E., Harris, J. B. & LaRocque, R. C. Antimicrobial-resistant bacteria in international travelers. *Curr. Opin. Infect. Dis.* **34**, 423–431 (2021).
9. Frost, I., Van Boeckel, T. P., Pires, J., Craig, J. & Laxminarayan, R. Global geographic trends in antimicrobial resistance: the role of international travel. *J. Travel Med.* **26**, taz036 (2019).
10. Conforti, S. *et al.* Strengthening Policy Relevance of Wastewater-Based Surveillance for Antimicrobial Resistance. *Environ. Sci. Technol.* **59**, 2339–2343 (2025).
11. Munk, P. *et al.* Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat. Commun.* **13**, 7251 (2022).
12. Prieto Riquelme, M. V. *et al.* Demonstrating a Comprehensive Wastewater-Based Surveillance Approach That Differentiates Globally Sourced Resistomes. *Environ. Sci. Technol.* **56**, 14982–14993 (2022).
13. Harrington, A. *et al.* Urban monitoring of antimicrobial resistance during a COVID-19 surge through wastewater surveillance. *Sci. Total Environ.* **853**, 158577 (2022).
14. Zhang, A. *et al.* β-lactam resistance genes in bacteriophage and bacterial DNA from

wastewater, river water, and irrigation water in Washington State. *Water Res.* **161**, 335–340 (2019).

15. Joseph, S. M., Battaglia, T., Maritz, J. M., Carlton, J. M. & Blaser, M. J. Longitudinal Comparison of Bacterial Diversity and Antibiotic Resistance Genes in New York City Sewage. *mSystems* **4**, e00327-19 (2019).

16. Cooper, L. N. *et al.* Socioeconomic Disparities and the Prevalence of Antimicrobial Resistance. *Clin. Infect. Dis.* **79**, 1346–1353 (2024).

17. See, I. *et al.* Socioeconomic Factors Explain Racial Disparities in Invasive Community-Associated Methicillin-Resistant Staphylococcus aureus Disease Rates. *Clin. Infect. Dis.* **64**, 597–604 (2017).

18. Andreatos, N., Shehadeh, F., Pliakos, E. E. & Mylonakis, E. The impact of antibiotic prescription rates on the incidence of MRSA bloodstream infections: A county-level, US-wide analysis. *Int. J. Antimicrob. Agents* **52**, 195–200 (2018).

19. Nadimpalli, M. L., Chan, C. W. & Doron, S. Antibiotic resistance: a call to action to prevent the next epidemic of inequality. *Nat. Med.* **27**, 187–188 (2021).

20. Arcilla, M. S. *et al.* Import and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. *Lancet Infect. Dis.* **17**, 78–85 (2017).

21. Sridhar, S. *et al.* Insights into global antimicrobial resistance dynamics through the sequencing of enteric bacteria from U.S. international travelers. Preprint at https://doi.org/10.1101/2025.01.27.635056 (2025).

22. Macias, E. P. & Morales, L. S. Crossing the Border for Health Care. *J. Health Care Poor Underserved* **12**, 77–87 (2001).

23. Corbett, K. *et al.* Appropriate antibiotic use: variation in knowledge and awareness by Hispanic ethnicity and language. *Prev. Med.* **40**, 162–169 (2005).

24. Hawkes, B. A. *et al.* Healthcare System Distrust and Non-Prescription Antibiotic Use: A

Cross-Sectional Survey of Adult Antibiotic Users. *Antibiotics* **12**, 79 (2023).

25. Zoorob, R., Grigoryan, L., Nash, S. & Trautner, B. W. Nonprescription Antimicrobial Use in a Primary Care Population in the United States. *Antimicrob. Agents Chemother.* **60**, 5527–5532 (2016).

26. Prieto Riquelme, M. V. *et al.* Demonstrating a Comprehensive Wastewater-Based Surveillance Approach That Differentiates Globally Sourced Resistomes. *Environ. Sci. Technol.* **56**, 14982–14993 (2022).

27. Caneschi, A., Bardhi, A., Barbarossa, A. & Zaghini, A. The Use of Antibiotics and Antimicrobial Resistance in Veterinary Medicine, a Complex Phenomenon: A Narrative Review. *Antibiotics* **12**, 487 (2023).

28. Mundkur, M. L. *et al.* Changes in Outpatient Use of Antibiotics by Adults in the United States, 2006–2015. *Drug Saf.* **41**, 1333–1342 (2018).

29. Grigoryan, L. *et al.* Use of Antibiotics Without a Prescription in the U.S. Population: A Scoping Review. *Ann. Intern. Med.* **171**, 257–263 (2019).

30. Munk, P. *et al.* Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat. Commun.* **13**, 7251 (2022).

31. Garner, E. *et al.* Metagenomic Profiling of Internationally Sourced Sewage Influents and Effluents Yields Insight into Selecting Targets for Antibiotic Resistance Monitoring. *Environ. Sci. Technol.* **58**, 16547–16559 (2024).

32. Brown, C. L. *et al.* Metagenomics disentangles epidemiological and microbial ecological associations between community antibiotic use and antibiotic resistance indicators measured in sewage. Preprint at https://doi.org/10.1101/2024.12.11.24318846 (2024).

33. Vikesland, P. *et al.* Can Wastewater-based Surveillance Relate Antimicrobial Use to Resistance Patterns? Metagenomic Analysis of Two Neighboring Treatment Plants. Preprint at https://doi.org/10.21203/rs.3.rs-4420801/v1 (2024).

34. Maestre-Carballa, L., Navarro-López, V. & Martinez-Garcia, M. City-scale monitoring of

antibiotic resistance genes by digital PCR and metagenomics. *Environ. Microbiome* **19**, 16 (2024).

35. Davis, B. C., Vikesland, P. J. & Pruden, A. Evaluating Quantitative Metagenomics for Environmental Monitoring of Antibiotic Resistance and Establishing Detection Limits. *Environ. Sci. Technol.* acs.est.4c08284 (2025) doi:10.1021/acs.est.4c08284.

36. Boehm, A. B. *et al.* Human viral nucleic acids concentrations in wastewater solids from Central and Coastal California USA. *Sci. Data* **10**, 396 (2023).

37. Huisman, J. S. *et al.* Wastewater-Based Estimation of the Effective Reproductive Number of SARS-CoV-2. *Environ. Health Perspect.* **130**, 057011 (2022).

38. Topol, A. High Throughput RNA Extraction and PCR Inhibitor Removal of Settled Solids for Wastewater Surveillance of SARS-CoV-2 RNA v1. Preprint at https://doi.org/10.17504/protocols.io.btyrnpv6 (2021).

39. Borchardt, M. A. *et al.* The Environmental Microbiology Minimum Information (EMMI) Guidelines: qPCR and dPCR Quality and Reporting for Environmental Microbiology. *Environ. Sci. Technol.* **55**, 10210–10223 (2021).

40. Esri Data and Maps. USA ZIP Code Boundaries. https://www.arcgis.com/home/item.html?id=d6f7ee6129e241cc9b6f75978e47128b (2023).

41. U.S. Centers for Disease Control and Prevention. Social Vulnerability Index 2022. https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html?CDC_AAref_Val=https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html.

42. Epic Systems Corporation. About Epic Cosmos. https://cosmos.epic.com/about/ (2025).

43. U.S. Census Bureau. 2022 5-Year American Community Survey (ACS). https://www.census.gov/programs-surveys/acs.

44. U.S. Census Bureau. 2020 Census Urban-Rural Classification. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

(2024).

45. Esri Data and Maps. USA Airports 1,000,000 or more. https://hub.arcgis.com/datasets/esri::usa-airports/about?layer=1.

46. US Department of Homeland Security. Homeland Infrastructure Foundation-Level Data. https://hifld-geoplatform.hub.arcgis.com/ (2024).

47. U.S. Department of Agriculture. 2022 National Agricultural Statistics Service Quick Stats. https://quickstats.nass.usda.gov/.

48. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063–1095 (2012).

49. Genuer, R., Poggi, J.-M. & Tuleau, C. Random Forests: some methodological insights. Preprint at https://doi.org/10.48550/ARXIV.0811.3619 (2008).

50. Kuhn, M. Building Predictive Models in *R* Using the **caret** Package. *J. Stat. Softw.* **28**, (2008).

Acknowledgments

We thank the participating wastewater treatment plants for their samples for the project. This work was supported by a gift to A.B.B. from the Sergey Brin Family Foundation. R.Y.L. was supported by National Institutes of Health (NIH) Institutional National Research Service Award T32 AI007502. U.S. antibiotic prescription rate data used in this study came from Epic Cosmos, a dataset created in collaboration with a community of Epic health systems representing more than 284 million patient records from over 1,500 hospitals and 36,000 clinics from all 50 states, D.C., Lebanon, and Saudi Arabia, though only U.S. data was used in this study. A.J.P. is a Chan Zuckerberg San Francisco Biohub Investigator.

Author contributions

A.J.P. and A.B.B. conceptualized the project. S.K., A.Z., E.M.G.C., D.D., R.Y.L., and C.M. acquired primary and secondary data in the study. S.K., A.Z. and E.M.G.C. performed the analysis. S.K., A.Z., and E.M.G.C. wrote the first draft. All authors reviewed and edited the final manuscript.