Estimation of country-specific tuberculosis antibiograms using genomic data

Avika Dixit, MBBS, MPH, MBI[1,2], Luca Freschi[2], PhD, Roger Vargas, Jr, PhD[2,3], Matthias I. Groeschel, MD, PhD[2], Sabira Tahseen, MD[4], SM Masud Alam, MBBS, MPH, MPhil[5], SM Mostofa Kamal, M. Phil[6], Alena Skrahina, MD, PhD, DSc[7], Ramon P. Basilio, MD[8], Dodge R. Lim, RN[8], Nazir A Ismail, MD[9], Maha R. Farhat, MD, MSc[2,10]

[1]Division of Infectious Diseases, Department of Pediatrics, Boston Children's Hospital, Boston MA
[2]Department of Biomedical Informatics and [3]Center for Computational Biomedicine, Harvard Medical School, Boston MA
[4]National Tuberculosis Control Programme, Islamabad, Pakistan
[5]National Tuberculosis Control Program, Directorate General of Health Services, Ministry of Health and Family Welfare, Dhaka, Bangladesh
[6]National Institute of Diseases of the Chest and Hospital, Dhaka, Bangladesh
[7]Republican Research and Practical Centre for Pulmonology and Tuberculosis, Dolginovski Trakt, Minsk, Belarus
[8]National Tuberculosis Reference Laboratory, Research Institute for Tropical Medicine, Muntinlupa City, Metro Manila, Philippines
[9]Global Tuberculosis Programme, World Health Organization
[10]Massachussetts General Hospital, Boston, MA

Abstract

Background: Global tuberculosis (TB) drug resistance (DR) surveillance is largely focused on the drug rifampicin. We leveraged public and surveillance *M. tuberculosis* (*Mtb)* whole genome sequencing (WGS) data, to generate more comprehensive country-level resistance prevalence estimates (antibiograms) using *in silico* resistance prediction.

Methods: We curated and quality-controlled *Mtb* WGS data**.** We used a validated random forest model to predict phenotypic resistance to twelve drugs and bias-corrected for model performance, outbreak sampling, and resistance oversampling. We validated our estimates using a national DR survey conducted in South Africa.

Results: *Mtb* isolates from 29 countries (n=19,149) met sequence quality criteria. Marginal genotypic resistance estimates overlapped with the South African DR survey for all drugs except isoniazid and second-line injectables that were underestimated (n=3,134); among multi-drug resistant (MDR) TB, estimates overlapped for pyrazinamide and the fluoroquinolones. Globally, mono-resistance to isoniazid was estimated at 10.9% (95% CI: 10.2-11.7%, n = 14,012. Mono-levofloxacin resistance rates were highest in South Asia (Pakistan 3.4% [0.1-11%], n=111 and India 2.8% [0.08-9.4%], n=114). Rates of resistance discordance between isoniazid and ethionamide were high with 74.4% (IQR: 64.5-79.7%) of isoniazid resistant isolates predicted to be ethionamide susceptible. The global susceptibility rate to pyrazinamide and levofloxacin among MDR was 15.1% (95% CI: 10.2-19.9%, n=3,964).

Conclusions: This is the first attempt at global *Mtb* antibiogram estimation. DR prevalence in *Mtb* can be reliably estimated using public WGS and phenotypic resistance prediction for key antibiotics. Our results raise concerns about the empiric use of short-course fluoroquinolone regimens for drug susceptible TB in South Asia and suggest that ethionamide is an under-utilized drug in MDR treatment.

Introduction

Tuberculosis (TB) and its causative agent *Mycobacterium tuberculosis* (*Mtb*) are a persistent global health threat resulting in more than 10 million incident cases and 1.5 million deaths annually[1]. TB was only recently overtaken as the top infectious disease killer by the coronavirus disease 2019 (COVID-19) in the year 2020. One of the major challenges in TB control is the emergence of antibiotic-resistant TB that is difficult and expensive to cure[2] with favorable treatment outcomes achieved in only 56% of cases[1,3]. Improved strategies to tackle resistant TB first require improved local estimates of resistance burden. National estimates currently reported by the World Health Organization (WHO) focus on rifampin resistance. These rely on either modeling, periodic surveys, or pooling rifampicin testing data, often derived from NAATs, including cartridge-based tests or LPAs, for countries that routinely test for rifampicin resistance[1]. Resistance estimates for the remaining agents are even more limited and still, largely rely on expensive and labor-intensive culture-based antibiotic susceptibility testing. Surveillance of resistance to other first- and second-line agents, *e.g.* pyrazinamide, fluoroquinolones, is needed to guide disease control and to project patient eligibility to standardized regimens, including newer fluoroquinolone-based regimens for antibiotic susceptible TB[4] and short-course regimens for multi-drug resistance[5].

Whole-genome sequencing (WGS) of clinical *M. tuberculosis* isolates is increasingly performed for research, surveillance and, clinical care and increasingly representative of high TB prevalence settings[6]. Between 2000 and 2010, 395 *Mtb* genomes were published in the National Center for Biotechnology Information's Sequence Read Archive (SRA). This number rose to 79,716 between 2011 and 2020. Major motivators for sequencing include the study of TB transmission/outbreaks[7] as well as genotypic surveillance of MDR or rifampin resistance in TB in both high[8] and low burden countries[9]. While these efforts involve oversampling of MDR or rifampin-resistant isolates, they are less likely to oversample higher-level resistance including pre-XDR and XDR TB, as this information is difficult to obtain pre-sampling due to the laboratory cost and biohazard, especially in high burden settings.

Enabled by an improved understanding of genetic resistance mechanisms, prediction from WGS is now a reliable approach to resistance diagnosis for the majority of *Mtb* antibiotics[10,11]. Several methods have been developed to predict resistance *in silico* from WGS data, these span simpler methods, like direct association, to machine learning[12–15]. Among the best-performing methods that have been validated across 12 different antibiotics is a random forest classifier (see **Methods** for performance details)[16]. We leveraged this model to comprehensively survey TB antibiotic resistance at the country level using public and surveillance WGS data. We outline an approach to correct DR and outbreak oversampling bias, and the model's imperfect sensitivity and specificity. We validate our WGS based estimates of drug resistance burden against national drug resistance survey data for South Africa. The results are accessible to the user using a point-and-click open web platform that can be refined over time as new WGS data and models become available.

Methods

Data curation:

We curated *Mtb* genomes from NCBI, PATRIC and, published literature[8,10,17–19] **(Supplementary Table 1)**. We also included genotype and resistance phenotype data from the WGS-based resistance survey from seven countries led by the WHO[8]. This survey performed cluster sampling of *Mtb* isolates from new and retreatment TB patients. The methods used for the curation of *Mtb* genomes and associated metadata, including geographic data, have been described previously[20]. We used metatools_ncbi (https://github.com/farhat-lab/metatools_ncbi) to download geographic location metadata from NCBI for each isolate. A summary table is available at (https://github.com/farhat-lab/resdata-ng/blob/master/metadata/summary_tables/geo_sampling.txt).

Phenotypic DST

Phenotypic DST data were curated from PATRIC and the published literature. Details of phenotypic testing methodology used by each study are shown in **Supplementary Table 1**. For studies that reported minimal inhibitory concentrations, we converted the results into a binary variable (indicating sensitive or resistant) using critical concentrations endorsed by the WHO[21].

Genomic analysis/variant calling:

We used a previously validated genomic analysis pipeline for *Mtb*[22] with modifications as described earlier[20]. Briefly, reads were trimmed using PRINSEQ[23] setting the average phred score threshold to 20. Raw read data was confirmed to belong to the *Mtb* complex using Kraken[24]. Isolates with <90% mapping were excluded. Reads were aligned to H37Rv (GenBank NC000962.3) reference genome using BWA MEM[25]. Duplicate reads were removed using PICARD[26]. We excluded any isolates with coverage of <95% at 10x or higher in known antibiotic resistance regions (*katG, inhA* & its promoter, *rpoB, embA, embB, embC* & *embB* promoter, *ethA, gyrA, gyrB, rrs, rpsL, gid, pncA, rpsA, eis* promoter). Variants were called using Pilon[27] that uses local assembly to increase indel (insertions and deletions) call accuracy. Variants were excluded if Pilon filter indicated low coverage.

Implementation of Random Forest (RF) Predictor:

We used a previously described RF model[12] for *in-silico* resistance prediction. Briefly, this model was trained using non-lineage marker nonsynonymous mutations in 18 loci adjusting for class imbalance using weights. The performance of the model has been re-validated recently on 20,408 isolates (sensitivity: rifampin 95.0% [95% CI: 94.4-95.6], isoniazid 91.8% [91.0-92.5] among rifampin-resistant and 81.1% [78.4-84.0] among rifampin susceptible, pyrazinamide 61.3% [59.2-63.4] and levofloxacin 80.2% [71.6-88.2])[16].

Estimation of resistance burden/antibiograms by country:

We developed a procedure to correct for the possible oversampling of outbreaks in our dataset. To exclude genetically similar isolates, we first calculated pairwise SNP distance across all isolates. Among the entire dataset of 24,015 isolates, 703,755 total SNPs were identified of

which 50,396 were further excluded because they either had low Empirical Base-level Recall (EBR) score[28], were located in mobile genetic element regions, or were missing in >10% of isolates (**Supplementary Figure 1**) with 653,359 SNP sites remaining. We excluded 1,416 isolates that had >=10% missing calls at these SNP sites and further excluded 15,771 SNPs where the minor allele didn't occur in any remaining isolates with 637,588 SNPs remaining among 22,599 total isolates (**Supplementary Figure 1**). We then calculated pairwise Euclidean SNP distances using a custom script. Using the R package `igraph`[29], we identified closely related isolates that had a genetic distance of less than or equal to 10 SNPs for each country and randomly selected one isolate from each group of isolates where each isolate was <= 10 SNP apart from the others. In groups of isolates where all isolates were not <=10 SNP apart, we excluded isolates with the highest relatedness iteratively until the least related isolates (>10 SNPs apart) remained. We compared resistance burden estimates with and without this outbreak correction.

We calculated the proportion of isolates resistant to each antibiotic, by country, using the number of isolates predicted as resistant divided by total isolates available for prediction for each country. We focus on four basic resistance estimates: (1) the marginal proportion of resistance among all TB isolates available, and the conditional proportion of resistance to a specific agent among (2) rifampin-susceptible isolates that we label as *mono-resistant*, (3) rifampin-resistant isolates, and (4) MDR isolates. In each case, we estimated the bias-corrected prevalence of antibiotic resistance (φ) using the genotypic prevalence (θ), using the sensitivity (*se*) and specificity (*sp*) of the RF model, as given by the formula below and described in Zignol et al.[8]:

$$\varphi = \frac{\theta + sp - 1}{se + sp - 1}$$

A distinction was made for the performance of the RF model in predicting isoniazid resistance among rifampin susceptible and rifampin-resistant isolates. We additionally assessed the performance of the RF model in predicting eligibility for the short course regimen defined as MDR isolates susceptible to both pyrazinamide and levofloxacin (or moxifloxacin where available using phenotypic testing). Uncertainty around each parameter was propagated using a Bayesian model using R library `rjags` to interface with JAGS 4.3.0[30]. To set up the model, *se* and *sp* were assumed to follow a Beta distribution with parameters obtained using the method of moments. The model was updated and re-fit 10,000 times. The posterior distribution of *φ* was randomly sampled to estimate the mean and standard deviation of the bias-corrected prevalence.

We also corrected for resistance oversampling among sequenced isolates. The formula below details this correction. We used rifampin resistance rates reported by the WHO from country-specific surveillance. We calculated a composite WHO rifampin resistance estimate using the proportion of rifampin resistance among new and retreatment cases reported by each country.

The probability *(P)* of resistance *(R)* to antibiotic *A* is given by:

$$P(A = R) = P(A = R|RR)P(RR) + P(A = R|RS)P(RS)$$

where $P(RR)$ is available from the WHO, $P(RS)$ is given by $1 - P(RR)$, and $P(A = R \mid RR)$, and $P(A = R|RS)$, i.e. mono-resistance to *A*, were the bias-corrected estimates of antibiotic resistance among rifampin-resistant and rifampin-susceptible isolates, respectively. We generated 95% confidence intervals (CI) by calculating the combined variance (Var) using the delta method.

Validation using national survey data:
We used the results of the South African national TB antibiotic resistance survey from 2012-14[8,31] to study residual bias in genotype-based estimates of drug resistance after correction for outbreak sampling. We extracted the overall resistance estimates (among both new and previously treated patients) for each antibiotic reported in the survey and compared these to marginal estimates generated by our method for all isolates and for MDR isolates separately.

Other analysis, data, and code availability:
Antibiograms were plotted for each country using `ggplot2`[32(p2)]. All the custom scripts used for this analysis were written using R 3.6.1 and Python 3.7.4 are available on GitHub at https://github.com/farhat-lab/tb_antibiogram.

Results

Data and estimation of antibiograms
We identified 22,599 isolates with data on country-of-origin that met sequence quality criteria (**Methods**). Average sequencing depth across resistance genes was 115X and 99% of bases in these regions were covered at ≥10X sequencing reads (**Methods**). There were 82 countries represented by at least one isolate and 32 countries by at least 100 isolates that met sequence quality criteria.

Phenotypic resistance data: Of the 22,599 isolates, 12,023 had phenotypic drug susceptibility testing (DST) results (**Figure 1**). The 12,023 isolates originated from 43 countries; 11,343 were tested for both isoniazid and rifampicin, and 2,856 (25% [95% CI: 24-26%]) were resistant to both, *i.e.* were MDR. Among the MDR isolates, 554 (19% [95% CI: 18-21%]) were resistant to at least one second-line injectable (capreomycin, amikacin, or kanamycin), 456 (16% [95% CI: 15-17%]) were resistant to at least one fluoroquinolone (ofloxacin, ciprofloxacin, levofloxacin, or moxifloxacin) and 259 (9% [95% CI: 8-10%]) were resistant to both. Among the 17 countries with at least 100 isolates with phenotypic DST, we computed the raw frequency of MDR in the sample. In comparison to the WHO reported MDR rates, 16 of the 17 countries had a higher MDR rate confirming the concern of overrepresentation of MDR in public genomic data.

Due to overrepresentation of MDR/rifampicin resistance, we assessed phenotypic resistance patterns strictly among rifampicin susceptible isolates (n=8,581 from 43 countries, median 30 isolates per country [IQR=3-225]). Isoniazid mono-resistance by phenotypic assay was seen in 9% (780/8,581) of isolates globally; Peru had the highest proportion at 33% (95% CI: 28-37%,

n=423) and none were isoniazid mono-resistant from China (95% CI: 0.0-8.2%, n=43). Among isolates susceptible to rifampin with DST to levofloxacin and/or moxifloxacin, 1.6% (95% CI: 1.1-2.2%, n=1,906 from eight countries) were resistant. Among the four countries with at least 100 rifampin susceptible isolates with DST to at least one fluoroquinolone, the highest proportion of fluoroquinolone mono-resistance was found in Bangladesh (3.9%, 95% CI: 2.3-6.2%, n=431, **Supplementary Table 2**). Genotype-based antibiograms, as detailed below, showed trends consistent with these phenotypic patterns even after bias correction. Genotype-based prediction allowed for estimation in a larger number of countries. Concordance between phenotypic DST and genotypic prediction was high for first-line drugs and lower for second-line drugs and consistent with published validation data[16] (**Supplementary Results**).

Genotypic antibiogram estimation: To generate 12-drug antibiograms, we followed a three-step procedure. We first filtered isolates that may represent *Mtb* outbreaks. We next applied a Bayesian correction for the imperfect sensitivity and specificity of the *in silico* resistance model. And lastly, to generate marginal antibiograms, we marginalized over rifampicin resistance categories using the WHO reported rate of rifampicin resistance for that country (**Methods**). As a substantial proportion of MDR-TB cases are related to recent transmission[20,33], and because outbreak investigation is one application of *Mtb* WGS resulting in oversampling of specific resistance genotypes, we applied an outbreak correction (**Methods**) to the 22,599 isolates that met sequence quality criteria. This led to the exclusion of 2,354 isolates, with 20,245 isolates from 78 countries remaining. The median percentage of isolates excluded from each country was 14.7% (IQR: 0.0-15.6%). Denmark, Argentina, and Djibouti were excluded as they had fewer than 100 isolates remaining after outbreak correction. We provide estimates with and without outbreak correction for the 29 countries represented by at least 100 sequenced isolates (median 342 isolates per country [IQR: 198-829]) in **Supplementary Table 3**.

Comparison with national antibiotic resistance survey data

We used the South African national TB antibiotic resistance survey from 2012-14[8,31] to study residual bias in genotype-based estimates of drug resistance after correction for outbreak sampling (**Figure 2**). Marginal resistance estimates overlapped for all drugs except for isoniazid and second-line injectables. For the latter two drug/classes, the rate estimates were lower using pooled public WGS data (n=3,134) than in the national resistance survey. Among MDR isolates (n=268), estimates using public WGS data overlapped estimates reported in the national DR survey for pyrazinamide, levofloxacin and para-amino salicylic acid, but were higher for other drugs including ethambutol, and second-line injectables (**Figure 2**).

Country-level estimates of antibiotic resistance

In additional to marginal antibiograms, we generated country-level bias-corrected estimates of mono-isoniazid and mono-levofloxacin resistance *i.e.*, only among rifampicin susceptible isolates, for the 26 countries with at least 50 rifampin susceptible isolates available for analysis (**Figure 3A-B**). We also estimated resistance to eleven drugs including pyrazinamide, and levofloxacin among MDR-TB for the 15 countries with at least 100 sequenced MDR-TB isolates (**Supplementary Figure 3**).

Marginal rates of resistance: The global rate of isoniazid marginal resistance was 11.6% (95% CI: 8.5-14.6%, n = 19,149) across 29 countries. For 10 of the 29 countries, >30% of isolates were estimated resistant to isoniazid; five of these were former Soviet Union countries (**Supplementary Figure 2**). The highest isoniazid resistance rate was estimated for Russia (66%, 95% CI: 45-87%, n=829), followed by Ukraine (58%, 95% CI: 39-76%, n=957). Isolates from the five former Soviet Union countries also had the highest pyrazinamide resistance rates (**Supplementary Table 3**). For countries outside of the former Soviet Union, the highest isoniazid resistance rate was measured in the Philippines (43%, 95% CI: 40-46%, n=181), Portugal (39%, 95%CI: 38-40%, n=100) and Peru (39%, 95% CI: 31-42%, n=1,521). The lowest isoniazid resistance rate was seen in South Africa (4.7%, 95% CI: 3.3-6.2%, n=3,134) and Japan (8%, 95% CI: 6-10%, n=368).

Marginal rates of ethionamide resistance also showed a wide range globally. On one extreme was the Republic of Moldova with a rate of 32% (95% CI: 22-42%, n=278) while the lowest rate was seen in the United Kingdom at 0.15% (95% CI: 0.05-0.30%, n=2,831). All countries (n=29) had higher rates of resistance to isoniazid compared to ethionamide with a median difference of 16.3% (IQR: 9.5-28.1%). Among isoniazid resistant isolates (n=6,090 from 29 countries, median 145 [IQR: 96-246] isolates per country), a median of 74.4% (IQR: 64.5-79.7%) were predicted to be ethionamide susceptible. Of the 4,827 total isolates that were predicted to be isoniazid resistant but ethionamide susceptible, 4,477 (92.7%) harbored antibiotic resistance conferring mutations in *katG* but not in *inhA*. Phenotypic DST was available for 565 of the 4,477 isolates and of these 80.2% tested resistant to isoniazid but susceptible to ethionamide.

Mono-resistance to isoniazid and fluoroquinolones: Twenty six of the 29 countries were represented by at least 50 rifampin susceptible isolates (median 237 [IQR:116-500] isolates per country). The global rate of isoniazid mono-resistance was estimated at 10.9% (95% CI: 10.2-11.7%, n = 14,012 across the 26 countries and explained most of the marginal rate of isoniazid resistance (11.6%, 95% CI: 8.5-14.6%, n=19,149) as noted above. By country (**Figure 3A**), we found evidence of over-estimation and under estimation at the two extremes of mono-resistance to isoniazid: highest in the Philippines (40%, 95% CI: 30-51%, n=119), and lowest in South Africa (1.2%, 95% CI: 0.5-2%, n=2,746). We verified high genotype and phenotype concordance for isoniazid in the Philippines and found evidence for over sampling of INH mono-resistance in public WGS data compared with the DR survey (Supplementary Results). For South Africa, isoniazid mono-resistance was underestimated due to the lack of resistance mutations in 60% of isolates; the remaining 40% of isoniazid mono-resistance was missed due to rare mutations in *katG* and the *ahpC* promoter, but not in the *fabG1-inhA* promoter (**Supplementary Tables 8 and 9**).

The global rate of mono-resistance to levofloxacin was estimated at 0.1% (95% CI: 0.003-0.3%, n = 14,012). There was considerable geographic variation: Pakistan and India had the highest rate of levofloxacin resistance, at 3.4% (0.1-11.1%) and 2.8% (0.1-9.4%) of 111 and 114 rifampicin susceptible isolates, respectively. In Bangladesh, which had a high prevalence of fluroquinolone mono-resistance based on phenotypic data, the bias corrected genotypic estimate was 0.7% (0.02-2.3%, n=454). Results by country are shown in **Figure 3B**. For India,

we compared the estimates to the national TB antibiotic resistance survey performed in 2014-2016 and in which the prevalence of levofloxacin resistance among new TB patients was 2.7% (95% CI: 2.2-3.4%)[34] (**Supplementary Figure 3**).

Resistance among MDR isolates: Fifteen countries had at least 100 MDR isolates after filtering of closely related isolates (n=3,964, median 179 isolates/country [IQR: 147.5-299.5]). We estimated MDR antibiograms for the 15 countries and across the whole sample, with and without outbreak correction (**Supplementary Figure 4A-B, Supplementary Table 4**). The global estimated rate of pyrazinamide resistance among MDR was 62.7% (95%CI: 59.3-66.1%, n=3,964); by country-level this rate ranged from 80% (95% CI: 74-86%, n=968) in Peru to 41% (95% CI: 29-53%, n=148) in Thailand. The global rate of levofloxacin resistance among MDR-TB was 8.6% (95% CI: 0.5-19.8%, n=3,964); by country this ranged from 48% in Japan (95% CI: 27%-69%, n=135) to 0.98% in the DRC (95% CI: 0.02-3.59%, n=144).

Susceptibility to antibiotics used in the MDR-TB Short Course Regimen

We estimated the bias-corrected proportion of MDR isolate with combined susceptibility to antibiotics used in the short course regimen (also known as “Bangladesh regimen”) by country. Specifically, we focused on two antibiotic classes, pyrazinamide and moxifloxacin/levofloxacin[35,36], as our approach predicted resistance to these drugs reliably in comparison with the South African DR survey. This approach provides a best-case scenario of feasibility of the use of short course regimen because: (1) the sensitivity and specificity of the RF model to identify eligibility was 91.1% and 59.5%, respectively (**Supplementary Table 5 and 6**), and (2) it ignores resistance to ethambutol and kanamycin which can be common among MDR-TB patients globally. For example, in our sample of 382 isolates that were phenotypically MDR and susceptible to pyrazinamide and levofloxacin/moxifloxacin, 64% were phenotypically tested resistant to ethambutol.

Phenotypic resistance only allowed us to explore estimates for Peru, Russia, and South Africa because these countries had at least 100 MDR isolates with phenotypic data to the two classes of antibiotics (pyrazinamide and moxifloxacin/levofloxacin). Phenotypic resistance to one or more of these drugs was 87% (n=124) in South Africa, 73% (n=733) in Peru, and 48% (n=295) in Russia.

WGS data allowed us to estimate feasibility across 15 countries. We measured an average global bias-corrected estimate, using susceptibility rate to both pyrazinamide and levofloxacin among MDR-TB, of 15.1% (95% CI: 10.2-19.9%, n=3,964). This combined rate was high for Democratic Republic of Congo (60.7% [95% CI: 45.6-75.1%, n=144]), the Netherlands (60.5% [46.4-73.8%, n=178]) and former Soviet Union countries had the lowest combined rates (e.g., Azerbaijan: 3.5% [0.1-11.3%, n=179] and Moldova: 3.8% [0.1-12.1%, n=163]) (**Figure 4**). For Bangladesh, where the regimen was originally developed, the combined rate was 42.5% (19.8-64.6%, n=69).

Access to antibiogram estimates: Genotypic estimates are available through a point-and-click web interface at https://gentb.hms.harvard.edu/maps/antibiogram/ to allow for quick reference

by clinicians and public health practitioners. Users can view the global distribution of overall resistance estimates for each drug and filter by resistance estimates among rifampin susceptible or rifampin resistant isolates. Country-level estimates of resistance rates are viewable by clicking on each country.

Discussion

Using a large *Mtb* genomic dataset, we estimate antibiotic resistance rates to 12 first-line and second-line antituberculosis agents across 29 countries with correction applied for oversampling of antibiotic-resistant isolates and outbreaks, as well as for genotypic model performance. We demonstrate the feasibility of this pathogen sequencing-based approach for resistance surveillance, and validate the model estimates against systematic national drug resistance survey data. This approach circumvents the major constraint in DR surveillance to-date which is limited by the access to culture-based DST. Our results reinforce that public WGS data is increasingly representative of TB in high prevalence settings, especially countries with a high burden of MDR. Overall, the antibiograms generated here provide key insights into resistance prevalence and co-resistance patterns globally, and have implications for TB management including empiric short regimen use for both drug susceptible and MDR-TB.

In recent clinical trials, a four-month fluoroquinolone and rifapentine-based treatment regimen for antibiotic susceptible TB was shown to be non-inferior to the current standard of care[4]. As this regimen may be soon rolled out for TB treatment, we studied the proportion of rifampicin susceptible isolates (as a proxy for rifapentine susceptibility) that harbored resistance to moxifloxacin or levofloxacin. This resistance would not be detectable by the widely used GeneXpert MTB/RIF that only detects rifampin resistance mutations[1]. Globally, the estimated prevalence of late-generation fluoroquinolone resistance was low at <1%, yet in countries we studied in South Asia the rate was 20-30 times higher. Our estimate of the prevalence was consistent with the prevalence among new TB cases reported by the National Drug Survey in India[34] and with estimates among rifampin susceptible isolates from Pakistan[36]. We speculate that the higher rate of fluoroquinolone resistance, may be related to dysregulated or over-the-counter use of fluoroquinolones in those countries[37,38]. But other factors including bacterial fitness of fluoroquinolone resistance and transmissibility of such isolates are yet to be explored[20]. Overall, our results highlight the need for comprehensive diagnostics that identify antibiotic resistance with a quick turnaround time. These will aid in the rapid identification of patients eligible for the newer fluoroquinolone-containing regimen for antibiotic susceptible TB or short course regimen for MDR-TB.

We found a high proportion of ethionamide susceptibility among isoniazid resistant isolates across the countries studied. The side-effect profile of ethionamide is relatively worse as compared to that of isoniazid (e.g. hepatotoxicity seen in up to 5% of patients as compared to up to 3% with isoniazid)[39] and it is typically reserved as a second-line agent[40,41]. Our results support the wider use and consideration of this agent in treatment of isoniazid mono-resistant or MDR-TB.

This study had several limitations. This included sampling bias of antibiotic-resistance in public TB WGS data. We recognized this limitation and designed a multistep bias correction procedure. There was, however, residual bias for resistance to several drugs in MDR isolates *e.g.*, ethambutol, and second-line drugs, as well as mono-resistance to isoniazid for countries at both extremes in resistance frequency. For isoniazid mono-resistance, we confirmed this bias was due to oversampling in available public WGS for the Philippines. Nevertheless, our corrected marginal resistance estimates, and estimates for pyrazinamide and fluoroquinolones among MDR-TB, overlapped consistently with national drug resistance survey data. Another limitation is the imperfect sensitivity of genotypic models for predicting resistance. We note that the RF models used in this study had higher sensitivity and specificity than direct association as reported for the recently released WHO catalogue of resistance mutations[42], but there are still notable gaps in sensitivity for mono-resistance to isoniazid, and certain drugs like ethambutol and second-line injectables. Our results support that mono-resistance to isoniazid is often caused by rare mutations that do not occur in MDR isolates, and hence training separate models for isoniazid mono-resistance may be necessary. Isoniazid remains an important agent in TB treatment, but second-line injectables are being phased out from clinical practice and perhaps surveillance of resistance to these agents may no longer be needed. The novel anti-tuberculosis agents including bedaquiline have recently become cornerstone agents in MDR-TB therapy where drug access allows. We were unable to generate estimates for bedaquiline, linezolid and delamanid in this study due to the lack of reliable genotypic prediction methods for these drugs. Recent reports do suggest very low rates of resistance to these agents, in part due to their recent introduction to clinical practice[43]. Another limitation is the lack of clinical metadata that did not allow us to estimate antibiograms separately for new and previously treated TB cases. Lastly, antibiogram estimation was necessarily limited to countries well represented in the isolate dataset and does not yet cover all high TB burden countries. We anticipate the wider adoption of *Mtb* sequencing for routine resistance diagnosis in high TB burden countries, championed by agencies such as UNITAID[44] and the Gates foundation[45] to address several of the aforementioned limitations in future WGS-based surveillance efforts.

In conclusion, we present an effort at global and comprehensive resistance rate estimation by repurposing public pathogen genomic sequence data and leveraging state-of-the-art resistance prediction models. We have made these data publicly accessible for use by clinicians and public health practitioners globally. Acknowledging their limitations, these estimates can assist geography-specific strategies for the control of TB and drug resistance. With the expansion of WGS for use in TB surveillance programs, the data available to generate these estimates is expected to grow and can be leveraged to allow for the monitoring of trends in resistance over time.

Acknowledgements

We would like to thank Anna Dean of the Global Tuberculosis Programme, World Health Organization, Geneva, Switzerland for her support in data collection and coordination of this project.

Funding

This work was funded by National Institutes of Health (R01 AI55765) and the Harvard Global Health Institute's Burke Fellowship (MRF). AD was supported by the Boston Children's Hospital Office of Faculty Development (the Basic and Translational Executive Committee, the Clinical and Translational Research Executive Committee Faculty Career Development Fellowship) and the Bushrod H Campbell and Adah F Hall Charity Fund (Charles A King Trust Postdoctoral Fellowship). RVJ was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303. MIG was supported by the German Research Foundation (GR5643/1-1). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Competing interest

The authors declare no competing interests

Author's contributions

AD and MRF conceived and designed the study. AD, LF, RVJ, MIG, and MRF wrote scripts for the data analysis. AD conducted the data analysis. ST, SMA, SMK, AS, RB, DL, NI provided data for analysis, and contributed with guidance and advice throughout the project. AD and MRF wrote the first version of the manuscript and the final manuscript contained contributions from all authors. The final manuscript was read and approved by all authors.

References:

1. *Global Tuberculosis Report 2020*. World Health Organization; 2020. Accessed October 24, 2020. http://www.who.int/tb/publications/global_report/en/

2. Centers for Disease Control and Prevention. Drug-Resistant TB. Published December 29, 2020. Accessed March 26, 2021. https://www.cdc.gov/tb/topic/drtb/default.htm

3. Dheda K, Gumbo T, Maartens G, et al. The epidemiology, pathogenesis, transmission, diagnosis, and management of multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *Lancet Respir Med*. 2017;5(4):291-360. doi:10.1016/S2213-2600(17)30079-6

4. Dorman SE, Nahid P, Kurbatova EV, et al. Four-Month Rifapentine Regimens with or without Moxifloxacin for Tuberculosis. *N Engl J Med*. 2021;384(18):1705-1718. doi:10.1056/NEJMoa2033400

5. Moodley R, Godec TR. Short-course treatment for multidrug-resistant tuberculosis: the STREAM trials. *Eur Respir Rev*. 2016;25(139):29-35. doi:10.1183/16000617.0080-2015

6. Starks AM, Avilés E, Cirillo DM, et al. Collaborative Effort for a Centralized Worldwide Tuberculosis Relational Sequencing Data Platform. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2015;61Suppl 3:S141-146. doi:10.1093/cid/civ610

7. Cabibbe AM, Walker TM, Niemann S, Cirillo DM. Whole genome sequencing of Mycobacterium tuberculosis. *Eur Respir J*. 2018;52(5). doi:10.1183/13993003.01163-2018

8. Zignol M, Cabibbe AM, Dean AS, et al. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. *Lancet Infect Dis*. 2018;18(6):675-683. doi:10.1016/S1473-3099(18)30073-2

9. Walker TM, Merker M, Kohl TA, Crook DW, Niemann S, Peto TEA. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. *Clin Microbiol Infect*. 2017;23(3):161-166. doi:10.1016/j.cmi.2016.10.014

10. CRyPTIC consortium , on behalf of 100,000 Genomes Project. DNA Sequencing Predicts 1st-Line Tuberculosis Drug Susceptibility Profiles. *N Engl J Med*. 2018;379(15):1403-1415. doi:10.1056/NEJMoa1800474

11. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM. Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. *Genome Med*. 2019;11. doi:10.1186/s13073-019-0660-8

12. Farhat MR, Sultana R, Iartchouk O, et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. *Am J Respir Crit Care Med*. Published online February 24, 2016. doi:10.1164/rccm.201510-2091OC

13. Yang Y, Niehaus KE, Walker TM, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinforma Oxf Engl*. 2018;34(10):1666-1671. doi:10.1093/bioinformatics/btx801

14. Kouchaki S, Yang Y, Walker TM, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinforma Oxf Engl*. 2019;35(13):2276-2282. doi:10.1093/bioinformatics/bty949

15. Chen ML, Doddi A, Royer J, et al. Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction. *EBioMedicine*. 2019;43:356-369. doi:10.1016/j.ebiom.2019.04.016

16. Gröschel MI, Owens M, Freschi L, et al. GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. *Genome Med*. 2021;13(1):138. doi:10.1186/s13073-021-00953-4

17. Allix-Beguec C, Arandjelovic I, Bi L, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N Engl J Med*. 2018;379(15):1403-1415. doi:10.1056/NEJMoa1800474

18. Phelan JE, Lim DR, Mitarai S, et al. Mycobacterium tuberculosis whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. *Sci Rep*. 2019;9(1):9305. doi:10.1038/s41598-019-45566-5

19. Davis JJ, Wattam AR, Aziz RK, et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res*. 2020;48(D1):D606-D612. doi:10.1093/nar/gkz943

20. Ektefaie Y, Dixit A, Freschi L, Farhat MR. Globally diverse Mycobacterium tuberculosis resistance acquisition: a retrospective geographical and temporal analysis of whole genome sequences. *Lancet Microbe*. 2021;2(3):e96-e104. doi:10.1016/S2666-5247(20)30195-6

21. World Health Organization. *Technical Report on Critical Concentrations for Drug Susceptibility Testing of Medicines Used in the Treatment of Drug-Resistant Tuberculosis*. World Health Organization; 2018. Accessed October 7, 2020. http://www.who.int/tb/publications/2018/WHO_technical_report_concentrations_TB_drug_susceptibility/en/

22. Ezewudo M, Borens A, Chiner-Oms Á, et al. Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. *Sci Rep*. 2018;8(1):15382. doi:10.1038/s41598-018-33731-1

23. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinforma Oxf Engl*. 2011;27(6):863-864. doi:10.1093/bioinformatics/btr026

24. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15:R46. doi:10.1186/gb-2014-15-3-r46

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

26. Picard Tools - By Broad Institute. Accessed April 27, 2018. http://broadinstitute.github.io/picard/

27. Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*. 2014;9(11). doi:10.1371/journal.pone.0112963

28. Marin M, Vargas R, Harris M, et al. *Genomic Sequence Characteristics and the Empiric Accuracy of Short-Read Sequencing*. Bioinformatics; 2021. doi:10.1101/2021.04.08.438862

29. Csárdi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Syst*. Published online 2006. Accessed February 13, 2017. http://wbldb.lievers.net/10011687.html

30. Plummer M. *Rjags: Bayesian Graphical Models Using MCMC*.; 2019. https://CRAN.R-project.org/package=rjags

31. National Institute for Communicable Diseases. South African tuberculosis drug resistance survey 2012–2014.

32. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. 2016. Springer International Publishing : Imprint: Springer; 2016. doi:10.1007/978-3-319-24277-4

33. Kendall EA, Fofana MO, Dowdy DW. Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. *Lancet Respir Med*. 2015;3(12):963-972. doi:10.1016/S2213-2600(15)00458-0

34. Ministry of Health and Family Welfare, Government of India. *Report of the First National Anti-Tuberculois Drug Resistance Survey, 2014-16*.; 2018. Accessed January 31, 2019. https://tbcindia.gov.in/WriteReadData/l892s/4187947827National%20Anti-TB%20Drug%20Resistance%20Survey.pdf

35. Willby M, Sikes RD, Malik S, Metchock B, Posey JE. Correlation between GyrA substitutions and ofloxacin, levofloxacin, and moxifloxacin cross-resistance in Mycobacterium tuberculosis. *Antimicrob Agents Chemother*. 2015;59(9):5427-5434. doi:10.1128/AAC.00662-15

36. Zignol M, Dean AS, Alikhanova N, et al. Population-based resistance of Mycobacterium tuberculosis isolates to pyrazinamide and fluoroquinolones: results from a multicountry surveillance project. *Lancet Infect Dis*. 2016;16(10):1185-1192. doi:10.1016/S1473-3099(16)30190-6

37. Shet A, Sundaresan S, Forsberg BC. Pharmacy-based dispensing of antimicrobial agents without prescription in India: appropriateness and cost burden in the private sector. *Antimicrob Resist Infect Control*. 2015;4. doi:10.1186/s13756-015-0098-8

38. Sarwar MR, Saqib A, Iftikhar S, Sadiq T. Antimicrobial use by WHO methodology at primary health care centers: a cross sectional study in Punjab, Pakistan. *BMC Infect Dis*. 2018;18(1):492. doi:10.1186/s12879-018-3407-z

39. Ethionamide. In: *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. National Institute of Diabetes and Digestive and Kidney Diseases; 2012. Accessed July 26, 2021. http://www.ncbi.nlm.nih.gov/books/NBK548025/

40. *The WHO Treatment Guidelines for Drug-Resistant Tuberculosis, 2016 Update*. World Health Organization; 2016. Accessed February 13, 2017. http://www.ncbi.nlm.nih.gov/books/NBK390455/

41. Nahid P, Mase SR, Migliori GB, et al. Treatment of Drug-Resistant Tuberculosis. An Official ATS/CDC/ERS/IDSA Clinical Practice Guideline. *Am J Respir Crit Care Med*. 2019;200(10):e93-e142. doi:10.1164/rccm.201909-1874ST

42. World Health Organization. *Catalogue of Mutations in Mycobacterium Tuberculosis Complex and Their Association with Drug Resistance*.; 2021. https://www.who.int/publications/i/item/9789240028173

43. Vargas R, Freschi L, Spitaleri A, et al. The role of epistasis in amikacin, kanamycin, bedaquiline, and clofazimine resistance in Mycobacterium tuberculosis complex. *Antimicrob Agents Chemother*. Published online August 30, 2021:AAC0116421. doi:10.1128/AAC.01164-21

44. Seq&Treat. FIND. Accessed July 26, 2021. https://www.finddx.org/at-risk-populations/seq-treat/

45. Makoni M. Africa's $100-million Pathogen Genomics Initiative. *Lancet Microbe*. 2020;1(8):e318. doi:10.1016/S2666-5247(20)30206-8

Figures and Tables:

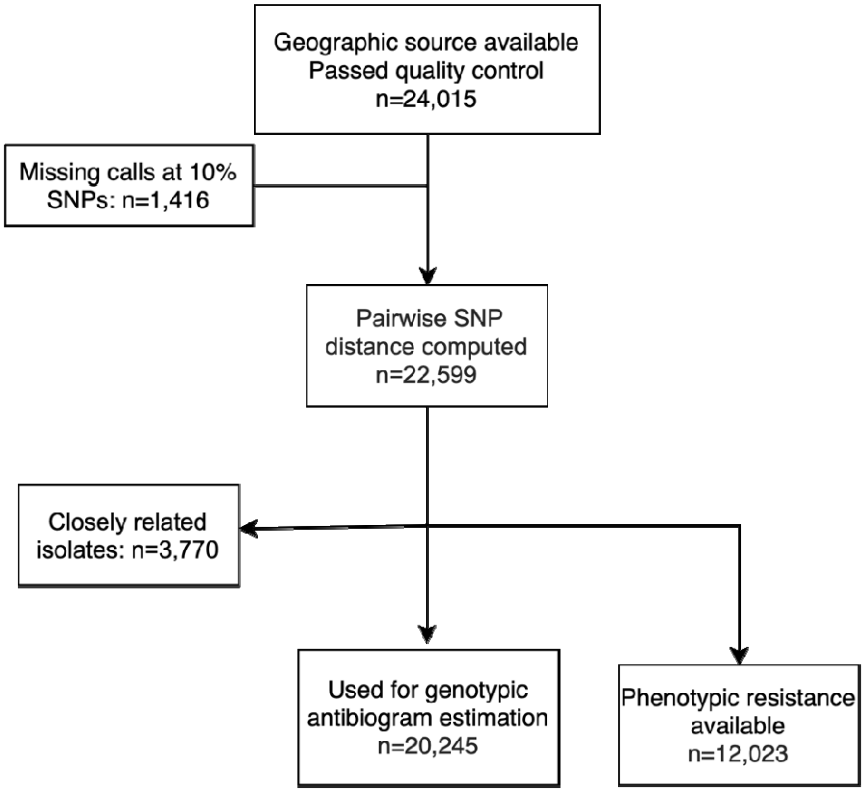

Figure 1: Isolates filtered and available for analysis. QC: Quality Control, SNP: Single Nucleotide Polymorphism. See **Supplementary Figure 1** for details on computation of pairwise SNP distance.

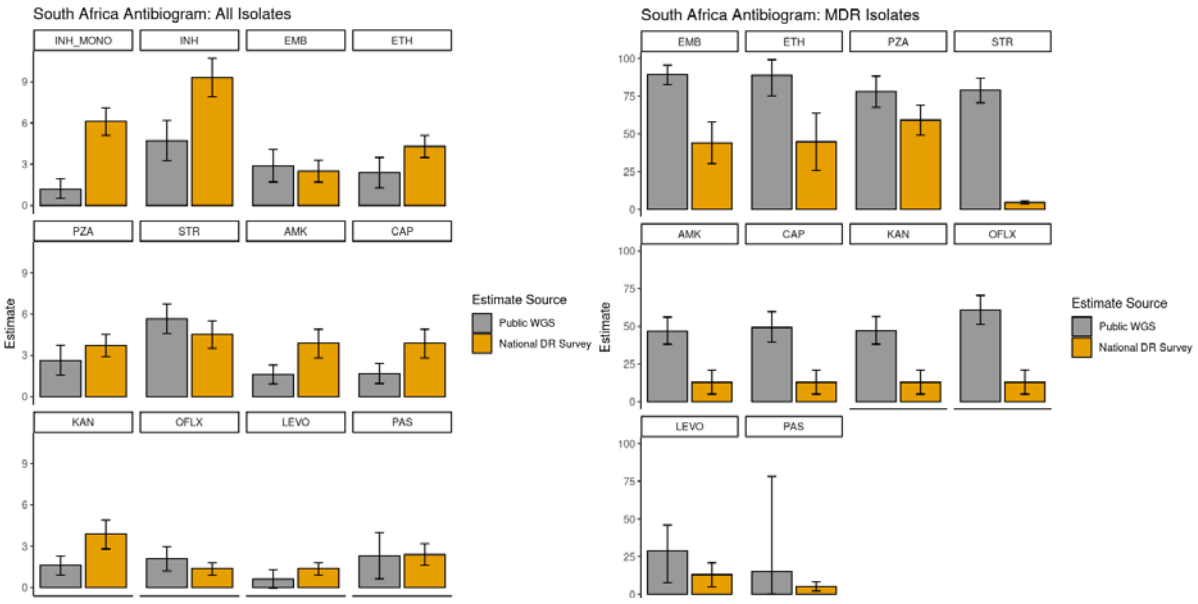

Figure 2: Validation of estimates for South Africa. Comparison of drug resistance estimates for South Africa from national drug resistance (DR) survey 2012-2014 to those calculated using public whole-genome sequences (WGS) in this study from A) all isolates (n=3,134) and, B) among multidrug-resistant (MDR) isolates (n=268), i.e., those resistant to both rifampin and isoniazid (INH). National DR survey reported second-line injectable resistance estimates were used to compare with amikacin (AMK), capreomycin (CAP), and kanamycin (KAN) estimates generated using public WGS. National DR survey reported ofloxacin (OFLX) resistance estimates were used to compare with OFLX and levofloxacin (LEVO) estimates generated using public WGS. EMB: ethambutol, ETH: ethionamide, INH_MONO: INH resistance in rifampicin susceptible isolates, PAS: para-amino salicylic acid, PZA: pyrazinamide, STR: streptomycin.

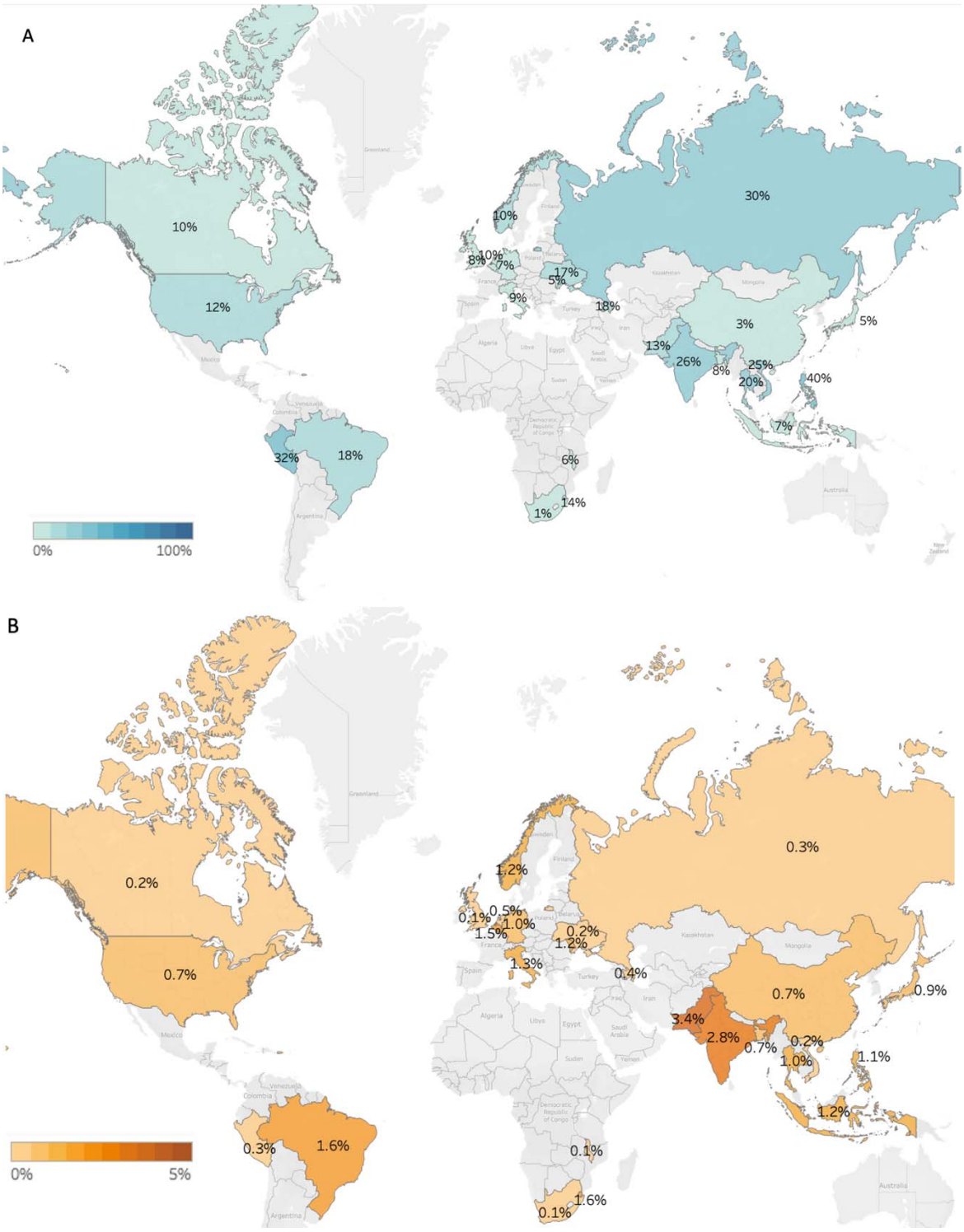

Figure 3: Bias-corrected estimates of (A) isoniazid mono-resistance and (B) levofloxacin mono-resistance in rifampin susceptible isolates. Only countries with at least 100 total isolates of which at least 50 were rifampin susceptible are shown.

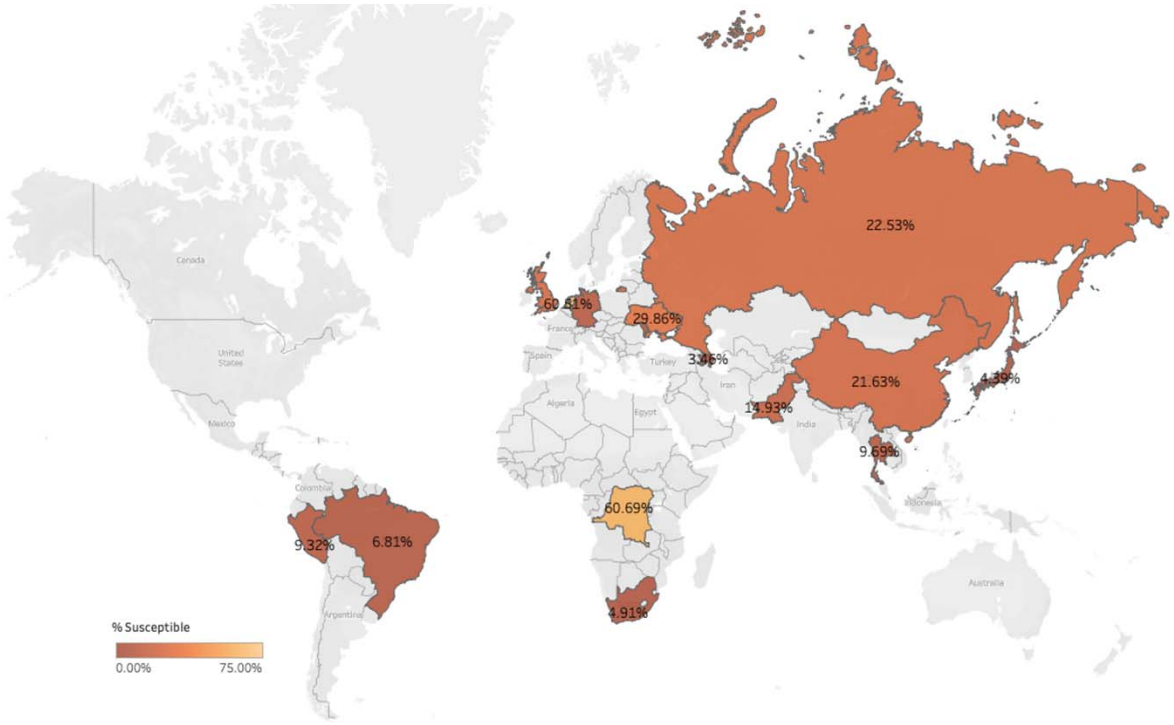

Figure 4: Bias-corrected estimates of susceptibility to antibiotics used for short course regimen for treatment of multidrug resistant tuberculosis (MDR-TB). Numbers shown for each country are bias-corrected percentage of MDR isolates that were sensitive to two of the antibiotics (pyrazinamide and levofloxacin) used in the short-course regimen. Only countries with at least 100 MDR isolates are shown.