Description of Additional Supplementary Files

File Name: Supplementary Data 1

Description: Numbers of isolates and variants dropped due to low-quality or unfixed variants and final counts used for training each of 135 regression models (9 models per drug). Unfixed variants have a within-isolate allele frequency in the range (0.25, 0.75]. The left bound is exclusive, and the right bound is inclusive. Isolates with a within-sample allele frequency of 0.25 are considered to not have the allele, and isolates with a within-sample allele frequency of 0.75 are considered missing with respect to that allele. A variant was dropped from the models if, after removing isolates with missingness, there were no remaining isolates with that variant present because regression models will assign an effect size of 0 to a variant that is absent from the dataset. Pooling combines variants into a single “LoF” variant, reducing the total number of variants. For some models, there was 1 or fewer such LoF variants. Pooling these variants did not affect the model, so the pooled models were not fit. The last two columns show the number and proportion of isolates in the intersection between each model dataset and the WHO dataset. The proportion is equal to the number of overlap isolates divided by the "# Final Isolates” column.

File Name: Supplementary Data 2

Description: Sample names and lineages for all 52,567 isolates used in this work. The “BioSample” column contains the public BioSample ID, if available; the “Sample_ID” column is the corresponding blinded sample ID used throughout this work; and the “Sample_Name” column should be used if no BioSample ID is available. Fast-lineage-caller was used to assign lineages according to 5 published schemes. The “Lineage” column is the primary lineage based on the Coll scheme, except for eight isolates that could not be classified by this scheme. These eight isolates are *M. canettii* by the Lipworth scheme, and the value in the “Lineage” column is “canettii.” When multiple lineages are called by the Coll scheme, all are reflected in the “Lineage” column.

File Name: Supplementary Data 3

Description: Regression gradings and statistics for 21,589 variants. Confidence intervals are 95% exact intervals based on the binomial distribution. For pooled LoF variants that were graded associated with R or not associated with R, all LoF component variants (frameshift, stop_gained, start_lost, and feature_ablation) in those genes were upgraded to Groups 2 or 4 in

the “REGRESSION + LOF UPGRADE” column. Descriptions of columns are in the second sheet.

File Name: Supplementary Data 4

Description: Mutations for which regression gradings supersede grading rules. There are 125 variants (79 Group 1-2, 46 Group 4-5) graded by regression but not by SOLO. These are upgraded using grading rules.

File Name: Supplementary Data 5

Description: MIC Model Results. Sheet 1: MIC regression results for 9,574 (drug, variant) pairs. All variants tested in the MIC models are given here. Sheet 2: 232 variants that were fit in the MIC models and have significant odds ratios in the WHO binary models (FDR p-value ≤ 0.05). There are 3 additional columns for the MIC coefficients, p-values, and FDR-adjusted p-values.

File Name: Supplementary Data 6

Description: Lineage Distribution of Regression Upgrades. Lineage and binary phenotype counts for 220 drug, variant pairs across 12 drugs graded by in Groups 1 and 2 by regression and in Uncertain by SOLO + grading rules.

File Name: Supplementary Data 7

Description: Source data for **Fig. 5.** Binary classification metrics across SOLO = Direct association method, SOLO = SOLO + additional grading rules, and regression. The SOLO statistics are from the published catalogue. Exact 95% confidence interval bounds and true positive, true negative, false positive, and false negative counts are listed.

File Name: Supplementary Data 8

Description: Binary model results for all 15 drugs with the major discrepancies katG_c.12A>G (isoniazid) and rrs_n.514A>C (capreomycin) removed. The only differences are for capreomycin. There is an additional row for isoniazid with the classification metrics if the 6 *ahpC* mutations that regression associates with resistance (Table 3) are removed. The name of this model row is “Regression – ahpC.”

File Name: Supplementary Data 9

Description: Source data for **Supplementary Fig. 6**. Binary classification metrics for AF thresholds of 75% and 25% for regression only, with exact 95% confidence interval bounds and true positive, true negative, false positive, and false negative counts.

File Name: Supplementary Data 10

Description: 2nd edition of the WHO Mutation Catalogue published in 2023, with both Tier 1 and 2 genes. Downloaded from https://github.com/GTB-tbsequencing/mutation-catalogue-2023/blob/main/Final%20Result%20Files

File Name: Supplementary Data 11

Description: Data contributors in the World Health Organization sequencing network.