AJBAR Vol. 5(5), October 2026: 115-139, ISSN: 2811-2881

Arid Zone Journal of Basic and Applied Research
Faculty of Science, Kashim Ibrahim University
Maiduguri, Nigeria

Journal homepage: https://www.azjournalbar.com

Review Article

Generalisability of Genomic Antimicrobial Resistance Prediction: External Validation, Geographic Representation and Clinical Readiness

Mustafa Alhaji Isa[1,4], Ibrahim Abubakar Mustapha[2] and Tijjani Isa Bulamarwa[3]
[1]Department of Biological Sciences, Kashim Ibrahim University, Maiduguri, Borno State, Nigeria
[2]Department of Islamic Studies, Kashim Ibrahim University, Maiduguri, Borno State, Nigeria
[3]Department of Physic, Kashim Ibrahim University, Maiduguri, Borno State, Nigeria
[4]Department of Microbiology, University of Maiduguri, Maiduguri, Nigeria
***Corresponding author's Email**: mustafaalhajiisa@gmail.com, doi.org/10.55639/607.020100154

ARTICLE INFO:

Keywords:
Antimicrobial Resistance, Whole-Genome Sequencing, Machine Learning, Susceptibility Prediction, External Validation.

ABSTRACT

Genomic prediction of antimicrobial resistance (AMR) offers an approach to earlier susceptibility assessment through bacterial genome sequencing, although the transportability of prediction models depends on genetic relatedness, clinical setting, antimicrobial exposure and laboratory practice. This review examines published primary evaluations and benchmark studies of genome-based phenotypic resistance prediction, focusing on internal versus independent validation, isolate provenance and the evidence needed for clinical implementation. Comparisons distinguish random partitions, lineage-aware evaluations, independent collections, temporal tests and laboratory-to-laboratory assessments. In a benchmark covering 11 bacterial species, 44 antimicrobials and 31,195 genomes, the proportion of experiments achieving macro F1 ≥0.9 was 64% under random splitting, 25% under phylogeny-aware splitting and 33% under homology-aware splitting. Individual investigations also describe variation across bacterial species, drugs, independent collections and bioinformatics pipelines. Susceptibility breakpoint definitions, the number of phenotypically resistant isolates, prediction abstention and calibration influence the clinical interpretation of these results. Geographic surveillance evidence further indicates an uneven distribution of publicly shared pathogen genomes, emphasising the importance of recording collection locations separately for development and validation cohorts. The evidence supports population-aware benchmarking, organism-drug-specific clinical error reporting, transparent treatment of uncertain calls and independent evaluation in the intended deployment population. An integrated framework is presented for assessing genomic predictors in relation to biological transportability, geographic representation and clinical utility.

Corresponding author: Mustafa Alhaji Isa, **Email**: mustafaalhajiisa@gmail.com
Department of Biological Sciences, Kashim Ibrahim University, Maiduguri, Borno State, Nigeria

INTRODUCTION

Antimicrobial resistance is among the leading causes of death worldwide. A systematic analysis of the global burden estimated that bacterial AMR was associated with 4.95 million deaths in 2019 and was directly attributable to 1.27 million deaths, with the

highest age-standardised mortality in western sub-Saharan Africa (Antimicrobial Resistance Collaborators, 2022). The clinical consequence of resistance is treatment failure, and the operational response depends on knowing, for a given isolate and a given drug, whether therapy is likely to work. For most of the history of clinical microbiology this knowledge has come from phenotypic antimicrobial susceptibility testing (AST), in which an organism is cultured and its growth is measured in the presence of defined drug concentrations.

Phenotypic AST is accurate and directly measures the quantity of interest, but it is slow. A susceptibility result typically follows one to three days of culture after a positive specimen, and slower-growing organisms such as *Mycobacterium tuberculosis* can require weeks. During this interval, clinicians prescribe empirically, which contributes both to under-treatment of resistant infections and to the overuse of broad-spectrum agents that drives further resistance. The appeal of inferring resistance directly from the genome is that the genetic determinants of resistance are present from the moment the organism is sampled, so a sequencing-based prediction could in principle be available as soon as sequence data are generated (Ellington et al., 2017).

Two broad strategies have been developed to predict resistance from genomes. The first is a rule-based or catalogue approach, in which the genome is searched for known resistance determinants, acquired genes or chromosomal mutations, and the presence of a determinant is mapped to a predicted resistant phenotype. Curated catalogues and tools such as those evaluated across this literature encode expert knowledge of validated mechanisms, and for some organism-drug combinations they achieve high concordance with phenotype (Kim et al., 2022; Mahfouz et al., 2020). The second strategy is statistical: a machine learning model is trained on genomes paired with measured phenotypes, and it learns features, whether individual variants, gene presence or absence, or short nucleotide words (k-mers), that discriminate resistant from susceptible isolates (Moradigaravand et al., 2018; Nguyen et al., 2018; Nguyen et al., 2019). Machine learning can, in principle, capture determinants that are not yet catalogued and can model minimum inhibitory concentration (MIC) as a graded outcome rather than a binary category (Nguyen et al., 2019).

The resources that underpin both strategies are now mature. Acquired resistance genes are identified against curated databases such as ResFinder (Zankari et al., 2012), the Comprehensive Antibiotic Resistance Database (Alcock et al., 2023) and the NCBI AMRFinder collection (Feldgarden et al., 2019), the last of which has been validated directly against genotype-phenotype correlations in defined isolate sets (Feldgarden et al., 2019). Chromosomal resistance is captured through organism-specific mutation catalogues, of which the World Health Organization catalogue for *Mycobacterium tuberculosis* is the most developed (World Health Organization, 2023). On the statistical side, reference-free methods that compare genomes through k-mers, such as the set-covering approach of Kover (Drouin et al., 2016) and the k-mer phenotype models of PhenotypeSeeker (Aun et al., 2018), learn predictive features without requiring a predefined determinant list. The distinction between the two families is not absolute: catalogue methods encode mechanism directly and are limited by the completeness of the catalogue, whereas statistical methods learn from data and are limited by the data they are given. The biology of the organism shapes how far either can succeed. Resistance that is mediated by a small number of well-characterised determinants in a largely clonal population is more tractable than resistance that is polygenic, regulated through expression, or mediated by combinations of acquired genes, efflux and chromosomal change in a genomically plastic population.

The two strategies are not mutually exclusive. Catalogue methods depend on the completeness of the underlying knowledge base, so they perform well where mechanisms are well characterised and poorly where they are not. Statistical methods depend on the data used to train them, so they perform well where the training data resemble the intended application and poorly where they do not. The central empirical question for both is therefore the same: under what conditions does a prediction learned or encoded in one setting remain reliable in another?

A recurring pattern in this field is the report of a high headline accuracy, obtained by splitting a single collection of isolates at random into a training portion and a test portion. Bacterial populations are strongly structured by descent: isolates fall into lineages, clones and outbreak clusters that share large portions of their genomes. When such a collection is divided at random, near-identical relatives of a test isolate are very likely to appear in the training set. A model can then achieve high apparent accuracy by recognising the lineage rather

than the causal determinant of resistance. This is a form of information leakage specific to genomic data, and it inflates estimates of how well a model will generalise to genuinely new isolates (Hicks et al., 2019; Hu et al., 2024; Kim et al., 2022).

The distinction that matters is therefore not between training and test data, but between a random partition of one development dataset and a test population that is independent in the dimensions along which the model will later be applied: different lineages, different hospitals, different countries and later time periods. A model that performs well on a random split has been shown only to interpolate within its own data distribution. A model that performs well across lineages, sites, regions and time has been shown to extrapolate, and extrapolation is what deployment requires (Hicks et al., 2019; Hu et al., 2024).

The published literature already contains broad reviews of artificial intelligence in microbiology, of genomic resistance prediction and of clinical implementation (Kim et al., 2022). The specific contribution needed is not another catalogue of algorithms but a structured comparison of internal performance against genuinely independent validation, with explicit attention to the populations that are poorly represented in the data. This review was designed to provide that comparison. Its objectives were:

1. to synthesise how genomic AMR prediction models perform under internal (random-split) evaluation compared with population-structure-aware, geographic, temporal, independent-dataset and prospective evaluation;
2. to characterise the geographic origin of the isolates used for model development and validation, distinguishing isolate origin from author affiliation, and to quantify the representation of low- and middle-income countries, including African settings;
3. to assess the evidence relevant to clinical readiness, including phenotype and breakpoint reporting, calibration, prediction uncertainty, false-susceptibility error and operational or prospective evaluation; and
4. to present a reusable evaluation framework and reporting checklist for future studies, and to identify the priority evidence gaps.

Virulence determinants are considered only where a study evaluates their contribution to resistance prediction. Therapeutic target discovery addresses a different question and is outside the scope of this review.

2. METHODS

This critical evidence review examines primary studies and methodological benchmarks of genome-based bacterial resistance prediction, together with relevant guidance on predictive-model evaluation and antimicrobial susceptibility testing (Collins et al., 2024; Moons et al., 2025; Page et al., 2021; Whiting et al., 2011). Evidence was organised by pathogen, antimicrobial, genomic feature, phenotypic reference method and validation design. The analysis distinguishes reported study outcomes from general methodological standards for evaluating the portability and clinical usefulness of genomic predictors.

2.1 REVIEW QUESTION

The review addressed a single question: how reliably do genomic resistance prediction models perform when applied to bacterial populations, hospitals, geographical regions and time periods different from those used for model development? The population of interest was bacterial isolates with paired genomic data and measured susceptibility; the exposure was a genomic prediction (rule-based or statistical); the comparator was measured phenotypic AST; and the outcomes were measures of discrimination, calibration and clinically relevant error, reported separately for internal and independent evaluation.

2.2 ELIGIBILITY CRITERIA

Studies were eligible if they predicted bacterial antimicrobial resistance phenotypes from genomic data (whole-genome sequence, assembled genome, gene or variant content, or k-mers) and evaluated those predictions against measured antimicrobial susceptibility. Both rule-based and machine learning approaches were eligible, as were studies of a single organism-drug pair and studies spanning many. Methodological and benchmarking studies that compared evaluation designs were eligible and were given particular weight because they speak directly to the review question.

- Excluded: studies predicting resistance gene presence without a phenotypic reference standard; studies predicting phenotype from non-genomic data only (for example mass spectrometry or clinical metadata alone) unless genomic data were also evaluated; narrative reviews, editorials and opinion pieces, which were retained only as context; and conference abstracts without sufficient methodological detail to chart.

- Where several publications reported on the same isolates and the same evaluation, the most complete report was charted and the others were noted, to avoid counting one dataset as independent evidence more than once.

2.3 INFORMATION SOURCES AND SEARCHES

Literature was identified through targeted searches linking bacterial resistance phenotypes, genome sequencing, machine learning, resistance catalogues and independent model evaluation, together with reference tracing from influential primary studies and benchmarks. The analysis considers studies with organism-level genomic predictors evaluated against phenotypic antimicrobial susceptibility testing. The evidence presented in Tables 1–4 is a purposively assembled collection of primary evaluations and benchmarks selected for their relevance to population structure, validation design, clinical errors and geographic provenance. The selection strategy permits a focused comparison of reported findings but does not provide an exhaustive census of publications.

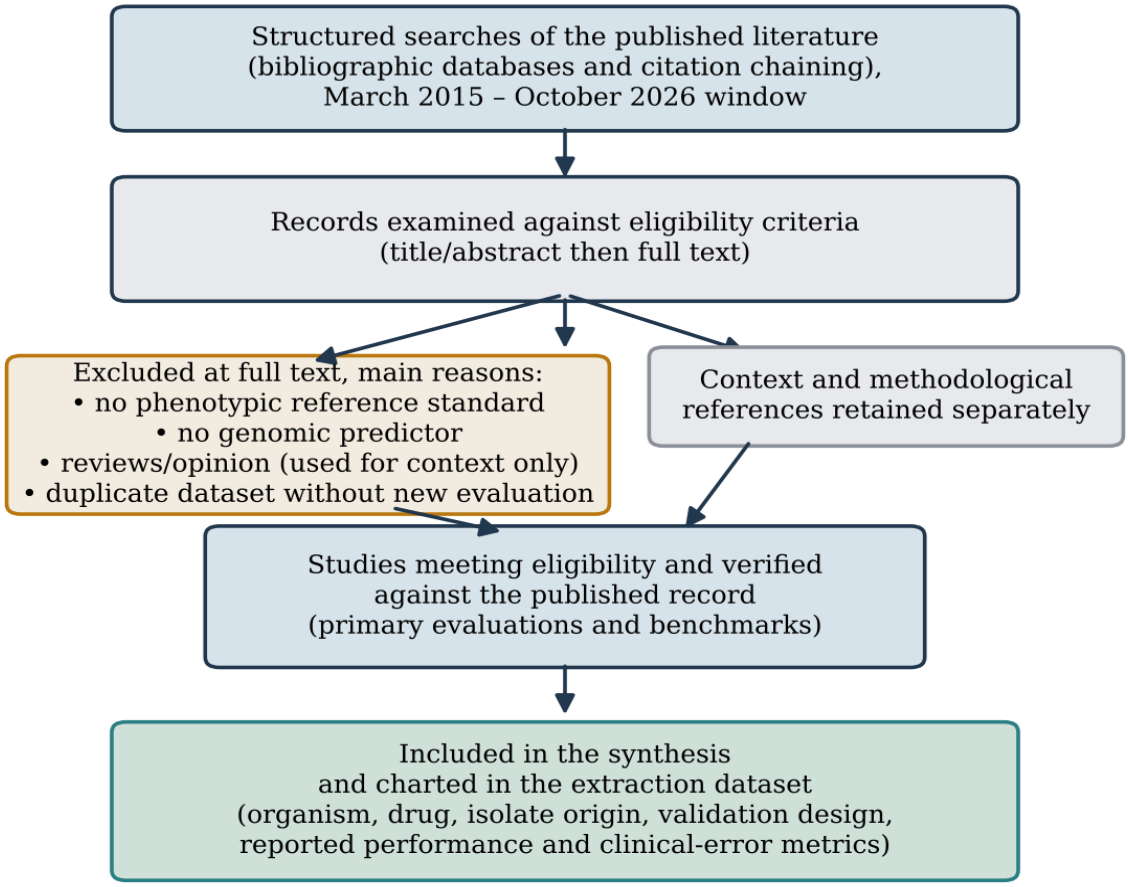


Figure 1. Assembly of the evidence base. Exact database export counts are to be finalised at the dual-screening stage (see Methods, Limitations).

Figure 1. Evidence identification and classification framework. Literature sources are organised by genomic prediction task, phenotypic reference standard, validation design and relevance to independent evaluation. The figure illustrates the evidence organisation process.

2.4 STUDY SELECTION AND DATA CHARTING

The evidence set was assembled through targeted identification and charting of published studies. Independent duplicate title–abstract screening, duplicate full-text eligibility assessment and duplicate extraction are not documented. The charting fields are: organism, antibiotic, number of

isolates, original collection country, AST method and breakpoint version, genomic features, algorithm, development dataset, training–test overlap, independent test setting, discrimination, calibration, uncertainty, and false susceptibility or false resistance with denominators. Entries marked "not fully stated" identify missing information in this working chart rather than definitive proof of nonreporting in the underlying publication.

Isolate origin was charted separately from author affiliation, because the location of a sequencing laboratory or a research group is not the location from which the isolates were collected. Where a study drew isolates from a public repository without stating their geographic provenance, the origin was charted as unreported rather than assumed.

2.5 RISK OF BIAS AND APPLICABILITY

Risk of bias and applicability were interpreted using the conceptual domains of PROBAST+AI, including data sources, predictors, reference outcomes and analysis (Moons et al., 2025). Particular attention was given to genomic relatedness across development and testing data, phenotype reference methods, resistance prevalence, calibration and the intended setting of application. These considerations inform the narrative assessment of individual study designs; the descriptive evidence-readiness matrix in Figure 5 summarises reporting dimensions rather than formal risk-of-bias judgements.

2.6 SYNTHESIS

Published estimates are retained at their original organism, antibiotic, outcome and validation setting. The quantitative values reproduced from the major benchmark represent its reported experiments, not independently recalculated effects across the evidence set. No meta-analysis or cross-study average decline was undertaken. Paired internal–external contrasts should be computed only when the same trained model, thresholds, susceptible/resistant definitions and comparable reference standards are available. Very major error rates use phenotypically resistant isolates as their denominator, whereas major error rates use phenotypically susceptible isolates; studies reporting alternative definitions must be shown separately. The proposed quantitative protocol is provided in Appendix A.

3. RESULTS

3.1 CHARACTERISTICS OF THE INCLUDED STUDIES

The selected primary evaluations cover Mycobacterium tuberculosis, Escherichia coli, Klebsiella pneumoniae, Salmonella, Pseudomonas aeruginosa, Staphylococcus aureus, Neisseria gonorrhoeae and Acinetobacter baumannii, together with multi-species benchmarks (Aytan-Aktug et al., 2020; Bradley et al., 2015; CRyPTIC Consortium and the 100,000 Genomes Project, 2018; Doyle et al., 2020; Hicks et al., 2019; Hu et al., 2024; Khaledi et al., 2020; Kim et al., 2020; Kouchaki et al., 2019; Mahfouz et al., 2020; Moradigaravand et al., 2018; Nguyen et al., 2018; Nguyen et al., 2019; Pataki et al., 2020; Ren et al., 2022; Walker et al., 2015; Walker et al., 2022). These investigations span catalogue-based predictors, sequence-feature models, quantitative MIC prediction and laboratory-to-laboratory comparisons. The studies provide organism- and drug-specific examples of development and validation across different settings. Table 1 summarises their reported characteristics, including isolate numbers and collection provenance where available.

Table 1. Characteristics of the verified included studies.

Study	Organism(s)	Drugs (n)	Isolates (n)	Isolate origin	Design
Walker 2015 (Walker et al., 2015)	*M. tuberculosis*	multiple	3,651	Multi-country	Derivation + validation set
Bradley 2015 (Bradley et al., 2015)	*S. aureus*; *M. tuberculosis*	12	2,995	UK, Germany, intl.	Derivation + validation set
CRyPTIC 2018 (CRyPTIC Consortium and the 100,000 Genomes Project, 2018)	*M. tuberculosis*	4	10,209	16 countries, 6 continents	Genotype–phenotype concordance

Nguyen 2018 (Nguyen et al., 2018)	*K. pneumoniae*	20	1,668	Clinical (USA)	10-fold cross-validation
Moradigaravand 2018 (Moradigaravand et al., 2018)	*E. coli*	11	1,936	UK (BSAC)	Nested CV, held-out
Kouchaki 2018 (Kouchaki et al., 2019)	*M. tuberculosis*	11	13,402	16 countries	Cross-validation
Hicks 2019 (Hicks et al., 2019)	*N. gonorrhoeae*; *K. pneumoniae*; *A. baumannii*	2	>1,400	Multiple datasets, intl.	Cross-dataset train/test
Nguyen 2019 (Nguyen et al., 2019)	Nontyphoidal *Salmonella*	15	5,278	USA (NARMS, 41 states)	CV + temporal/geographic subsets
Aytan-Aktug 2020 (Aytan-Aktug et al., 2020)	4 species	multiple	7,116	Public repositories	Species-independent models
Khaledi 2020 (Khaledi et al., 2020)	*P. aeruginosa*	4	414	Germany / Europe	Cross-validation
Kim 2020 (VAMPr) (Kim et al., 2020)	9 species	29	3,393	Public repositories	Nested CV + external clinical sets
Mahfouz 2020 (Mahfouz et al., 2020)	5 Gram-negative species	multiple	2,587	PATRIC / NDARO	Database–phenotype concordance
Pataki 2020 (Pataki et al., 2020)	*E. coli*	1 (CIP)	704	Multi-country	Cross-validation
Doyle 2020 (Doyle et al., 2020)	6 Gram-negative species	4	7 isolates / 10 datasets	UK, multi-laboratory	Inter-laboratory concordance
Ren 2022 (Ren et al., 2022)	*E. coli*	4	987 + 1,509	Germany + public dataset	CV + external public dataset
Walker 2022 (Walker et al., 2022)	*M. tuberculosis*	13	38,215	45 countries	Catalogue grading
Hu 2024 (benchmark) (Hu et al., 2024)	11 species	44	31,195	PATRIC (global)	Random + phylogeny + homology-aware

BSAC, British Society for Antimicrobial Chemotherapy; NARMS, National Antimicrobial Resistance Monitoring System (USA); NDARO, National Database of Antibiotic Resistant Organisms; CV, cross-validation; CIP, ciprofloxacin; intl., international. Isolate counts are as reported; where a study combined development and validation sets, the combined total is shown.

3.2 GENOMIC INPUTS, PHENOTYPE REFERENCE STANDARDS AND REPORTING

The genomic inputs fell into three families. Catalogue and rule-based methods searched the genome for known determinants and mapped them to a phenotype (Kim et al., 2022; Mahfouz et al., 2020; Walker et al., 2022). Variant-based machine learning used the presence or absence of specific mutations and genes, often restricted to curated

resistance loci, as model features (Kim et al., 2020; Pataki et al., 2020). Alignment-free machine learning used k-mers, short nucleotide words counted across the whole genome, which capture both known and uncharacterised signal at the cost of interpretability (Moradigaravand et al., 2018; Nguyen et al., 2019; Ren et al., 2022). Several studies modelled the MIC directly as a graded outcome rather than reducing it to a binary susceptible-or-resistant call (Nguyen et al., 2018; Nguyen et al., 2019).

The quality of a prediction can be no better than the quality of the phenotype against which it is trained and judged. Phenotypic AST is itself variable: a measured MIC is reproducible only to within approximately one two-fold dilution, and categorical calls near a breakpoint are unstable. Studies that modelled MIC and reported agreement within one dilution were implicitly acknowledging this measurement floor (Nguyen et al., 2018; Nguyen et al., 2019). The breakpoint standard and its version matter because category boundaries defined by the Clinical and Laboratory Standards Institute (Clinical and Laboratory Standards Institute, 2026) and the European Committee on Antimicrobial Susceptibility Testing (European Committee on Antimicrobial Susceptibility Testing, 2026) change over time, and a prediction judged against one standard may be scored differently against another; one systematic assessment found that recalculating categories with a more recent breakpoint set changed balanced accuracy only modestly on average, but this cannot be assumed for every organism-drug pair (Mahfouz et al., 2020). The handling of intermediate results and of missing phenotypes was reported inconsistently. Table 2 charts the genomic inputs and the phenotype reporting for the included studies, and shows that breakpoint standard and version, and the handling of intermediate categories, were frequently not stated.

Table 2. Genomic inputs and phenotype reporting in the included studies**.**

Study	**Genomic input / features**	**Susceptibility reference**	**Breakpoint standard / version**	**Intermediate / missing handling**
Walker 2015 (Walker et al., 2015)	Catalogue of resistance mutations (9+ genes)	Culture-based DST	Not fully stated	Unclassified isolates reported (10.8%)
Bradley 2015 (Bradley et al., 2015)	Catalogue (Mykrobe); de Bruijn graph	Culture-based AST/DST	Not fully stated	Not fully stated
CRyPTIC 2018 (CRyPTIC Consortium and the 100,000 Genomes Project, 2018)	Mutation catalogue, 9 genes	Phenotypic DST	Not fully stated	Predictions withheld for unknown mutations
Nguyen 2018 (Nguyen et al., 2018)	k-mers (MIC regression)	Broth microdilution MIC	CLSI (MIC agreement used)	MIC modelled directly
Moradigaravand 2018 (Moradigaravand et al., 2018)	Gene presence/absence, SNPs, pop. structure	Categorical susceptibility	Not fully stated	Not fully stated
Nguyen 2019 (Nguyen et al., 2019)	k-mers (MIC regression)	Broth microdilution MIC (NARMS)	MIC agreement ±1 dilution	MIC modelled directly
Hicks 2019 (Hicks et al., 2019)	Unitigs / variants	MIC, categorical	CLSI and EUCAST compared	Near-breakpoint isolates examined

Khaledi 2020 (Khaledi et al., 2020)	Gene presence, variants, expression	Categorical susceptibility	Not fully stated	Not fully stated
Aytan-Aktug 2020 (Aytan-Aktug et al., 2020)	ResFinder/PointFinder features	Categorical (public)	Not fully stated	Not fully stated
Kim 2020 (VAMPr) (Kim et al., 2020)	Ortholog-based variant genotypes	Categorical (public)	Not fully stated	Not fully stated
Mahfouz 2020 (Mahfouz et al., 2020)	CARD / ResFinder markers	Categorical (PATRIC/NDARO)	Recalculated with 2019 breakpoints	Class imbalance handled (balanced accuracy)
Pataki 2020 (Pataki et al., 2020)	gyrA/parC mutations, qnr genes	Ciprofloxacin MIC	MIC agreement (2- and 4-fold)	MIC modelled directly
Doyle 2020 (Doyle et al., 2020)	AMR gene detection (multiple tools)	Phenotypic AST	Not fully stated	Coverage thresholds examined
Ren 2022 (Ren et al., 2022)	k-mers, FCGR encodings	Categorical susceptibility	Not fully stated	Down-sampling for class balance
Walker 2022 (Walker et al., 2022)	Mutation catalogue, 13 drugs	Phenotypic DST (pooled)	WHO/technical criteria	Graded by confidence; low-prevalence drugs flagged
Hu 2024 (benchmark) (Hu et al., 2024)	Multiple (k-mer, catalogue, ML)	Categorical (PATRIC)	Not fully stated	Negative-class metrics reported separately

DST, drug susceptibility testing; AST, antimicrobial susceptibility testing; SNP, single-nucleotide polymorphism; FCGR, frequency chaos game representation; CLSI, Clinical and Laboratory Standards Institute; EUCAST, European Committee on Antimicrobial Susceptibility Testing. "Not fully stated" indicates that the field was not reported in sufficient detail to chart unambiguously.

3.3 VALIDATION DESIGNS AND WHAT EACH ESTABLISHES

The evaluation designs encountered in the literature form a hierarchy of increasing stringency, and each establishes a different and larger claim about generalisation (Figure 2). A random split of a single collection establishes only that a model can interpolate within its own data distribution; because near-identical relatives are shared between training and test portions, it is the weakest guard against the population-structure leakage described in Section 1.3. A lineage-aware or phylogeny-aware split, in which whole clades or clusters are assigned entirely to training or to testing, establishes that a model can transfer across genetic backgrounds. An independent-hospital or independent-collection design establishes transfer across sites and laboratory protocols. A geographic split establishes transfer across the distinct bacterial populations that circulate in different regions. A temporal split, training on earlier years and testing on later ones, establishes stability as resistance mechanisms and their frequencies evolve. A prospective or clinical design, in which the prediction is made before the phenotype is known and is then compared with the subsequent laboratory result, is the only design that establishes operational readiness.

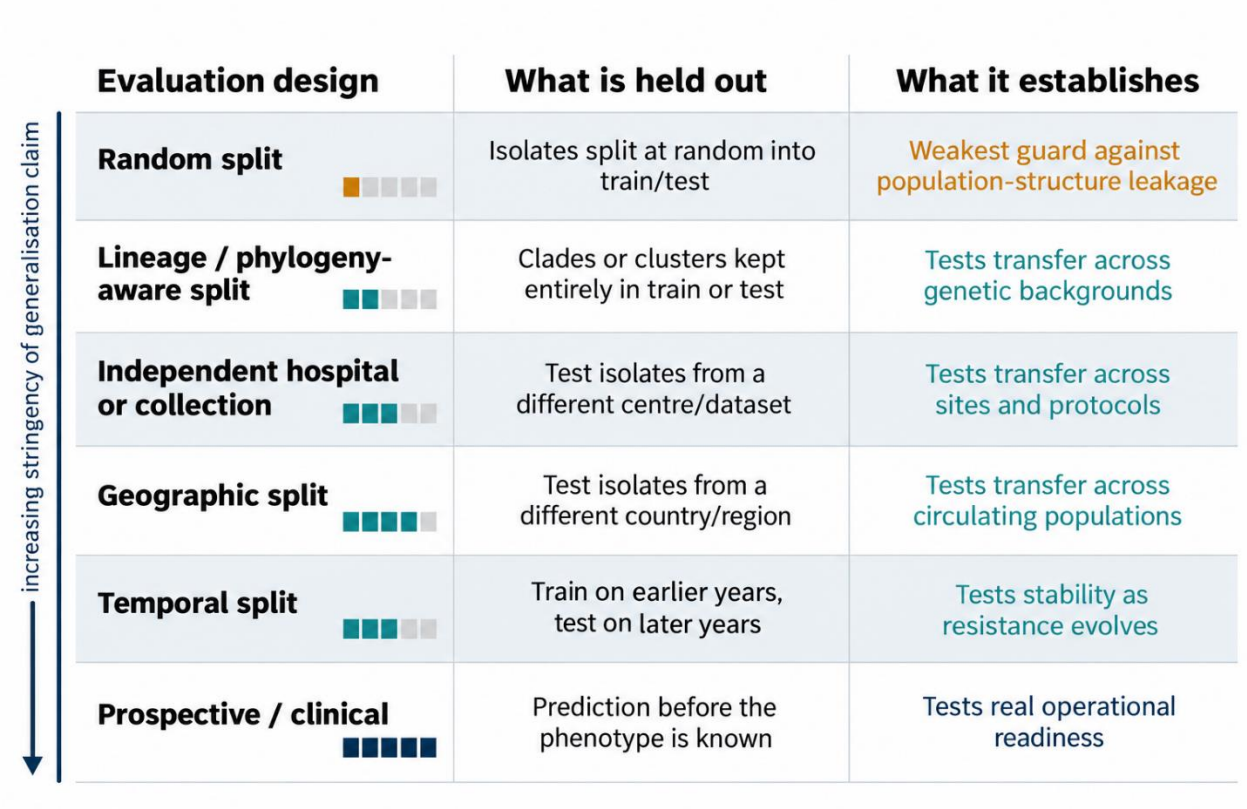


Figure 2. A framework of validation designs for genomic resistance prediction. Each design holds out a different dimension of variation and therefore establishes a progressively larger claim about generalisation. Shaded bars indicate the relative stringency of the generalisation claim each design supports.

The included studies occupied the lower and middle rungs of this hierarchy far more often than the upper ones. Random and simple cross-validation designs were the most common (Aytan-Aktug et al., 2020; Khaledi et al., 2020; Kim et al., 2020; Moradigaravand et al., 2018; Nguyen et al., 2018; Ren et al., 2022). Population-structure-aware evaluation was the explicit focus of the multi-method benchmark (Hu et al., 2024) and of the gonococcal cross-dataset study (Hicks et al., 2019), and was partially addressed through nested or phylogenetically informed procedures in several others. Temporal evaluation was reported in the *Salmonella* MIC study, where models were trained on earlier years and tested on later ones (Nguyen et al., 2019). Independent-collection evaluation was reported where a model trained on one dataset was tested on a separately assembled dataset (Hicks et al., 2019; Ren et al., 2022) and where predictions were compared across laboratories (Doyle et al., 2020). Genuinely prospective clinical evaluation, in which a genomic prediction was issued before the phenotype and then audited, was essentially absent from the model-development literature and was identified by the clinical-perspective review as a central missing piece (Kim et al., 2022). Table 3 charts the model families, comparators and evaluation designs of the included studies.

Table 3. Model families, comparators and evaluation designs.

Study	Model family	Baseline / comparator	Split design	Independent test population

Walker 2015 (Walker et al., 2015)	Catalogue mapping	Expert rules	Derivation + validation	Separate validation set
Bradley 2015 (Bradley et al., 2015)	Catalogue (Mykrobe)	Phenotype	Derivation + validation	Separate validation set
CRyPTIC 2018 (CRyPTIC Consortium and the 100,000 Genomes Project, 2018)	Catalogue	Phenotype	Concordance analysis	Multi-country pooled
Nguyen 2018 (Nguyen et al., 2018)	XGBoost (MIC)	None stated	10-fold CV	No (within-collection)
Moradigaravand 2018 (Moradigaravand et al., 2018)	Gradient-boosted trees; others	Population-structure-only model	Nested CV, held-out	No (within-collection)
Kouchaki 2018 (Kouchaki et al., 2019)	Logistic regression, GBT, +PCA/NMF	Direct association	Cross-validation	No (within-collection)
Hicks 2019 (Hicks et al., 2019)	Set covering machine; random forest	Single-dataset training	Cross-dataset train/test	Yes (independent datasets)
Nguyen 2019 (Nguyen et al., 2019)	XGBoost (MIC)	Combined vs per-antibiotic	CV + temporal/source/geographic	Partial (temporal, source subsets)
Aytan-Aktug 2020 (Aytan-Aktug et al., 2020)	Neural networks	Species-specific models	Cross-validation	Partial (cross-species)
Khaledi 2020 (Khaledi et al., 2020)	ML classifiers (multi-omic)	Single data type	Cross-validation	No (within-collection)
Kim 2020 (VAMPr) (Kim et al., 2020)	Explainable ML (variant)	Known mechanisms	Nested CV + external	Yes (external clinical sets)
Mahfouz 2020 (Mahfouz et al., 2020)	Rule-based (CARD, ResFinder)	Between databases	Repository concordance	Yes (external repositories)
Pataki 2020 (Pataki et al., 2020)	Linear model (4 predictors)	Full ML model	Cross-validation	Multi-country collection
Doyle 2020 (Doyle et al., 2020)	Multiple tools (per laboratory)	Reference PCR / phenotype	Inter-laboratory	Yes (blinded, cross-laboratory)
Ren 2022 (Ren et al., 2022)	RF, SVM, LR, CNN	Between model families	CV + external dataset	Yes (external public dataset)

Walker 2022 (Walker et al., 2022)	Catalogue grading	Expert review	Catalogue construction	Pooled multi-country
Hu 2024 (benchmark) (Hu et al., 2024)	4 ML tools + ResFinder	Majority-class baseline; ResFinder	Random + phylogeny + homology-aware	Yes (structure-aware, leave-one-species-out)

GBT, gradient-boosted trees; PCA, principal component analysis; NMF, non-negative matrix factorisation; RF, random forest; SVM, support vector machine; LR, logistic regression; CNN, convolutional neural network.

3.4 BENCHMARK AND REFERENCE RESOURCES

Three large resources anchor the quantitative evidence and are summarised in Table 3.1. The first is a multi-method benchmark that evaluated four machine learning tools and a rule-based catalogue method across 78 species-antibiotic datasets, covering 11 species, 44 antimicrobials and 31,195 genomes, under random, phylogeny-aware and homology-aware partitions; it is the only resource that manipulates the relatedness between training and test isolates directly, and its released folds and dataset identifiers support reproducible cross-study comparison (Hu et al., 2024). The second is an internationally assembled tuberculosis compendium that paired the genomes of 12,289 *M. tuberculosis* isolates from 23 countries with quantitative MICs to 13 drugs measured on a harmonised assay, which reduces one major source of laboratory heterogeneity while leaving differences in circulating lineages and resistance prevalence to be examined (CRyPTIC Consortium, 2022). The third is the WHO catalogue of resistance-associated mutations, drawn from tens of thousands of isolates across many countries, which provides a mechanism-based interpretive baseline whose per-drug and per-region coverage is itself uneven (World Health Organization, 2023). These resources are infrastructure for evaluation rather than evidence of any single model's external performance, and the sections that follow interpret the model studies against them.

Table 3.1. Principal benchmark and reference resources.

Resource	Scale	Role	Interpretive caution
Multi-method benchmark (Hu et al., 2024)	11 species; 44 drugs; 78 species-drug datasets; 31,195 genomes	Compares methods across random, phylogeny-aware and homology-aware splits	Manipulates relatedness directly, but is not a substitute for geographic or prospective evaluation
Tuberculosis compendium (CRyPTIC Consortium, 2022)	12,289 *M. tuberculosis* isolates; 23 countries; 13 drugs	Harmonised genome-MIC resource for evaluating transfer	Standardised phenotype does not remove lineage or prevalence differences
WHO mutation catalogue (World Health Organization, 2023)	Tens of thousands of isolates; many countries; 13 drugs	Mechanism-based interpretive baseline	Geographic breadth does not guarantee uniform per-drug or per-region coverage

3.5 GENERALISATION PERFORMANCE: INTERNAL VERSUS INDEPENDENT EVALUATION

The clearest evidence on the central question comes from studies that evaluated the same models under more than one design. The most comprehensive is a multi-method benchmark that applied four machine learning tools and a rule-based catalogue method to 78 species-antibiotic datasets, covering 11 species, 44 antimicrobials and 31,195 genomes, under random, phylogeny-aware and homology-aware partitions (Hu et al.,

2024). The effect of the evaluation design was large and consistent. Machine learning methods reached a macro F1 of at least 0.9 in 64% of experiments under random splits, but in only 25% under phylogeny-aware splits and 33% under homology-aware splits (Figure 4). At the less demanding threshold of a macro F1 of at least 0.8, the proportions were 81%, 50% and 60% respectively. The same ordering held for negative-class precision and F1, the metrics most relevant to avoiding false-susceptibility errors.
.

The authors reported that machine learning performance declined as test strains became more evolutionarily divergent from the training data, and that in a leave-one-species-out setting all methods performed significantly worse than their single-species counterparts. These are not failures of particular algorithms; they are the expected consequence of evaluating interpolation and extrapolation separately, and they quantify how much of the headline performance in this field is attributable to population-structure leakage

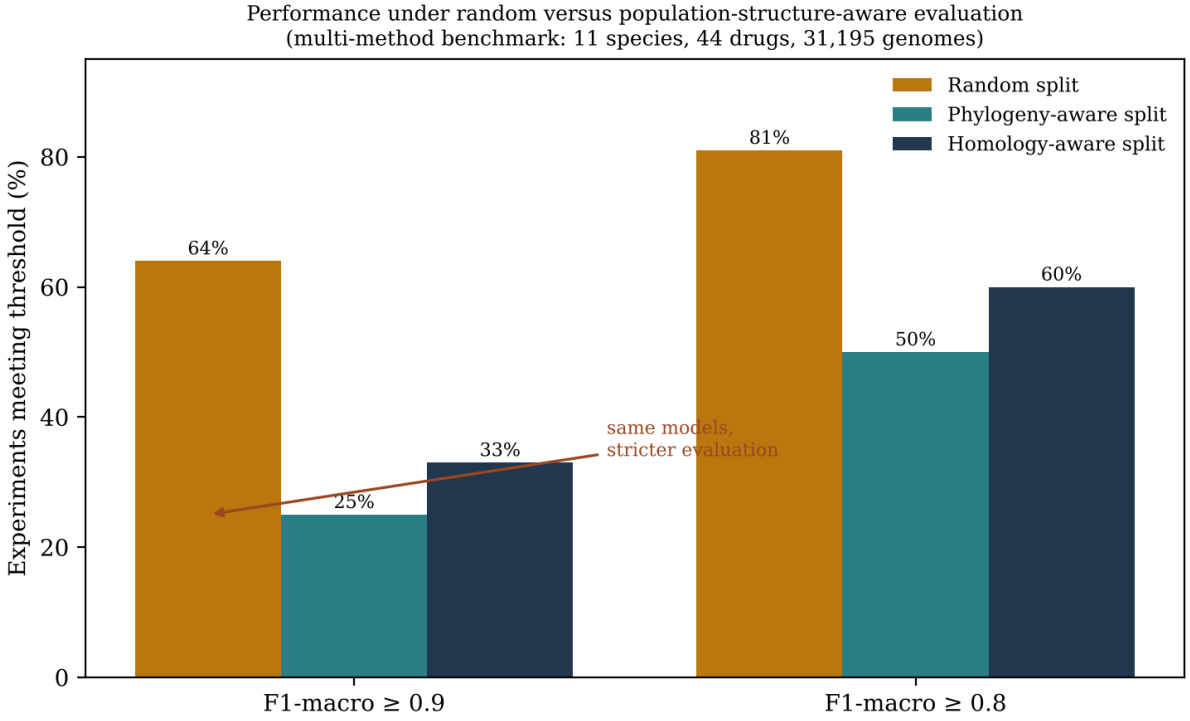


Figure 4. The same models, evaluated under random versus population-structure-aware partitions, in a multi-method benchmark of 11 species, 44 drugs and 31,195 genomes. The proportion of experiments reaching each macro F1 threshold falls substantially when near-identical relatives are prevented from spanning the training and test sets. Values from reference (Hu et al., 2024).

Organism identity strongly modified the size of this effect. In the benchmark, *Campylobacter jejuni* and *Enterococcus faecium* remained accurate across all three designs, whereas *E. coli*, *M. tuberculosis*, *K. pneumoniae* and, most markedly, *P. aeruginosa* performed less well, with the latter showing the lowest median performance under random splits before any structure-aware penalty was applied (Hu et al., 2024). This gradient tracks genome biology. Organisms whose resistance is dominated by a small number of well-characterised chromosomal mechanisms, and whose populations are comparatively clonal, present a prediction problem that both catalogues and models solve well and that transfers across collections. Organisms with large accessory genomes, frequent horizontal gene transfer and resistance mediated by combinations of acquired genes, efflux and regulatory change present a problem that is harder to learn and harder to transfer. The strong performance repeatedly reported for *M. tuberculosis* (Bradley et al., 2015; CRyPTIC Consortium and the 100,000 Genomes Project, 2018; Kouchaki et al., 2019; Walker et al., 2015; Walker et al., 2022) and for *Salmonella* MIC

prediction (Nguyen et al., 2019) should be read in this light: these are, in genomic terms, among the more tractable targets, and their success does not generalise automatically to the Gram-negative pathogens that dominate the clinical resistance burden.

Individual studies that included an independent-dataset step reinforced the benchmark's conclusion. In the gonococcal cross-dataset analysis, models trained on an aggregate of datasets did not generally outperform models trained on the test dataset's own isolates, and excluding the test dataset from training significantly reduced performance for almost every dataset examined, with the behaviour of each model tracking the ratio of non-susceptible to susceptible isolates in the data (Hicks et al., 2019). In an *E. coli* study, a convolutional neural network did not consistently outperform a random forest, and model performance differed between the laboratory's own collection and an external public dataset (Ren et al., 2022). The inter-laboratory study found that when the same sequence data were analysed independently by nine groups, pairwise agreement between genotypic resistance predictions was 79% across antibiotics and samples, and only 2% of pairwise gene-set comparisons were identical; against phenotype, overall sensitivity was 76% and specificity 50%, improving markedly only when low-coverage samples were excluded (Doyle et al., 2020). Reproducibility of the prediction itself, independent of any population-structure question, is therefore not assured.

3.6 GEOGRAPHIC REPRESENTATION OF TRAINING AND VALIDATION ISOLATES

Training data provenance is central to transferability. The present curated examples include single-country studies, multi-country tuberculosis collections and public repositories with varying metadata completeness. They cannot support a numerical estimate of the proportion of all AMR prediction publications using African isolates or the assertion that no such studies exist. The external amr.watch analysis of public genomes is useful background evidence but must not be conflated with the geographic distribution of model-training isolates (David et al., 2025). Each included study in a future full review should be assigned isolate-origin categories separately for development and external evaluation, including “unknown origin”; multiple-country datasets should be weighted by the number of isolates only when disaggregated counts are available.

A separate genomic-surveillance analysis documented marked geographical concentration of publicly shared genomes of priority pathogens (David et al., 2025). That finding identifies a plausible representativeness challenge, but genome-repository composition is not a direct count of the isolates used in prediction-model evaluation. Accordingly, any figure based on the surveillance analysis should be labelled as external context; a quantitative evidence map specific to model development and external validation requires extraction of accession-level isolate locations, their denominators and the fraction of unknown origins.

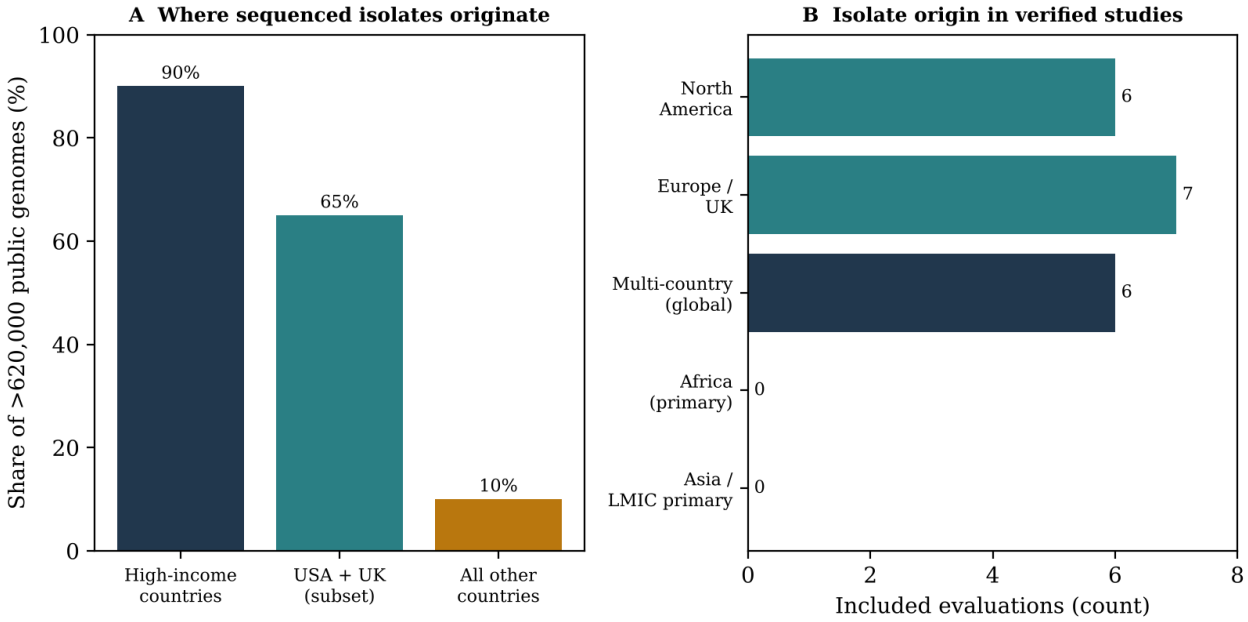


Figure 3. Illustrative geographic context: panel A reproduces reported proportions from an external public-genome surveillance study (David et al., 2025); panel B categorizes selected publications using stated study descriptions, not original isolate-weighted country counts. Categories can overlap, and the chart must not be interpreted as the geographic composition of all genomic prediction studies.

The consequence is not merely a question of equity, although it is also that. A model trained on isolates from one region learns the resistance mechanisms and their frequencies as they occur in that region. Where a different region carries different determinants, a different balance of clones, or a different prevalence of resistance, both the discrimination and the calibration of the model can degrade, and the degradation will be invisible to any evaluation conducted only within the original region. The under-representation of low- and middle-income settings is therefore simultaneously an equity gap and a validity gap, and it cannot be closed by analysis alone; it requires the generation and sharing of sequence-and-phenotype data from the under-represented settings themselves (David et al., 2025).

Isolate-weighted geographic comparisons depend on cohort-level country information, collection year, sequence accession and data use for model training, tuning or evaluation. Distinguishing the source of each isolate from the location of the submitting laboratory allows geographic coverage to be interpreted in relation to the actual bacterial populations represented. Isolates with unreported provenance form a separate category, and denominators distinguish individual isolates from studies, cohorts and isolate–drug evaluations.

Training cohorts and external validation cohorts provide distinct views of geographic representation. Comparisons across countries are most informative when isolate numbers, sampling periods and provenance completeness are available, while amr.watch provides a complementary global context for publicly shared genomes (David et al., 2025).

3.7 CLINICAL RELEVANCE: CALIBRATION, ERROR ASYMMETRY AND OPERATIONAL EVALUATION

For a prediction to be clinically useful it must not only discriminate but also be calibrated, be accompanied by a usable statement of uncertainty, and fail in the safer direction. The asymmetry of errors is central. A very major error, in which a resistant isolate is predicted susceptible, may lead to ineffective therapy and is the more dangerous failure; a major error, in which a susceptible isolate is predicted resistant, leads to unnecessarily broad therapy and is less dangerous but still consequential for stewardship. Regulatory expectations for diagnostic AST place tight limits on the very major error rate in particular.

Against these expectations the evidence was uneven. In the *Salmonella* MIC study, overall agreement within one two-fold dilution was high at 95.2%, and agreement was maintained across collection years, contamination sources, geographic states and serovars; very major error rates met regulatory criteria for several agents but exceeded them for others, and the authors attributed the latter to a scarcity of resistant isolates for those drugs (Nguyen et al., 2019). This dependence of the very major error rate on the number of resistant training examples is a recurring theme: where resistance is rare in the data, the model has little opportunity to learn it, and false-susceptibility errors rise precisely for the determinants that a clinician would most need to detect. In the systematic assessment of marker databases, false-susceptibility (very major) error rates were comparatively low but false-resistance (major) error rates were high, at 42.7% for one database and 25.1% for another, so that neither met the performance profile required of a clinical diagnostic without further development (Mahfouz et al., 2020). The inter-laboratory study, which most closely resembled operational conditions, reported specificity of only 50% against phenotype across the blinded analyses before quality filtering (Doyle et al., 2020).

Calibration and uncertainty were reported rarely. Few studies presented calibration curves or probabilistic outputs that would let a clinician weigh a prediction against its confidence, and few implemented a formal abstention mechanism, although the catalogue-based tuberculosis work did withhold predictions where isolates carried mutations of unknown significance, which is a principled form of abstention (CRyPTIC Consortium and the 100,000 Genomes Project, 2018; Walker et al., 2015). Prospective evaluation, in which predictions are issued before the phenotype and audited against it, was essentially absent. The clinical-perspective review drew the same conclusion from a wider reading of the field and located the central limitation not in discrimination but in the absence of calibrated, uncertainty-aware, prospectively evaluated evidence (Kim et al., 2022); the EUCAST subcommittee reached a concordant position, judging that for most species the evidence was insufficient to support the use of sequence-inferred susceptibility to guide clinical decisions, and recommending standardised benchmarks, quality metrics and a curated reference database before that step is taken (Ellington et al., 2017). Table 4 charts the discrimination and clinical-error evidence for the included studies.

Table 4. Discrimination and clinically relevant error in the included studies.

Study	Context	Discrimination (key metric)	Independent / structure-aware result	Error asymmetry / notes
Walker 2015 (Walker et al., 2015)	Mtb, validation set	Sens 92.3%, spec 98.4%	Validation set	10.8% unclassified (abstention)
Bradley 2015 (Bradley et al., 2015)	Sa; Mtb	Sa sens 99.3%/spec 99.5%; Mtb sens 82.6%/spec 98.5%	Validation set	Lower Mtb sensitivity
CRyPTIC 2018 (CRyPTIC Consortium and the 100,000 Genomes Project, 2018)	Mtb, 4 first-line drugs	INH sens 97.1%/spec 99.0%; RIF 97.5%/98.8%	Multi-country pooled	97.9% of predicted-pansusceptible correct
Nguyen 2018 (Nguyen et al., 2018)	Kp, 20 drugs (MIC)	92% within ±1 dilution	Within-collection	≥90% for 15/20 drugs

Moradigaravand 2018 (Moradigaravand et al., 2018)	Ec, 11 drugs	Mean accuracy 0.91	Within-collection	Pop.-structure-only model 0.79
Kouchaki 2018 (Kouchaki et al., 2019)	Mtb, 11 drugs	AUC gains vs direct association	Within-collection	Second-line F1 improved with dimension reduction
Hicks 2019 (Hicks et al., 2019)	Ng (CIP, AZM); Kp; Ab	CIP bACC ≥93%; AZM 57–94%	Cross-dataset: performance drops excluding test dataset	Sens/spec ratio tracks class balance
Nguyen 2019 (Nguyen et al., 2019)	*Salmonella*, 15 drugs (MIC)	95.2% within ±1 dilution	Temporal 86–92%; geographic/source ~95%	VME meets criteria for 7 drugs; rises where R rare
Khaledi 2020 (Khaledi et al., 2020)	Pa, 4 drugs (multi-omic)	Sens/PPV 0.8–0.9+	Within-collection	Expression improved 3/4 drugs
Kim 2020 (VAMPr) (Kim et al., 2020)	9 species, 29 drugs	Mean accuracy 91.1%	External clinical sets	Interpretable; recovers known mechanisms
Mahfouz 2020 (Mahfouz et al., 2020)	5 Gram-neg species	bACC 0.66 (ResFinder), 0.52 (CARD)	External repositories	ME 25–43%; VME 1–4%
Pataki 2020 (Pataki et al., 2020)	Ec, ciprofloxacin (MIC)	65% within 2-fold; 93% within 4-fold	Multi-country collection	4-predictor model as good as full ML
Doyle 2020 (Doyle et al., 2020)	6 Gram-neg species, 4 drugs	Pairwise consensus 79%	Blinded inter-laboratory	Sens 76%/spec 50%; 98% sens if >30× depth
Ren 2022 (Ren et al., 2022)	Ec, 4 drugs	RF AUC up to 0.96	External public dataset: performance differs	CNN not consistently best
Walker 2022 (Walker et al., 2022)	Mtb, 13 drugs	Sens >80% (5 drugs); spec >95% (most)	Pooled multi-country	Few mutations for new drugs (low R prevalence)
Hu 2024 (benchmark) (Hu et al., 2024)	11 species, 44 drugs	F1≥0.9 in 64% (random)	25% (phylogeny-aware); 33% (homology-aware)	Decline with evolutionary divergence

Sens, sensitivity; spec, specificity; PPV, positive predictive value; bACC, balanced accuracy; AUC, area under the receiver operating characteristic curve; VME, very major error (false susceptibility); ME, major error (false resistance); INH, isoniazid; RIF, rifampicin; CIP, ciprofloxacin; AZM, azithromycin; Mtb, M. tuberculosis; Sa, S. aureus; Ec, E. coli; Kp, K. pneumoniae; Ng, N. gonorrhoeae; Ab, A. baumannii; Pa, P. aeruginosa; R, resistant.

Clinical error is defined at the organism–drug level. With phenotypic resistance designated as the positive outcome, a very major error occurs when a resistant isolate is predicted susceptible, whereas a major error occurs when a susceptible isolate is predicted resistant. The corresponding rates use phenotypically resistant and phenotypically susceptible isolates, respectively, as denominators. No-call decisions, excluded isolates and uncertainty intervals influence their

clinical interpretation (Ellington et al., 2017; Mahfouz et al., 2020; Nguyen et al., 2019).

Probabilistic outputs add information beyond binary discrimination when they are accompanied by calibration curves, calibration intercepts and slopes, or appropriate scoring rules. Changes in resistance prevalence between training and independent evaluation populations may alter predictive values and calibration. Selective prediction can additionally be characterised through the fraction of isolates receiving a definitive result and the error rate within that fraction (Collins et al., 2024; Kim et al., 2022).

A quantitative comparison of diagnostic performance is most interpretable within a defined organism–drug pair, using compatible AST reference standards, independent cohorts and complete two-by-two outcome counts. Paired evaluation of a frozen predictor under internal and independent testing allows performance changes to be separated from changes in algorithms or decision thresholds. In the studies assembled here, differences in evaluation design, outcome definition and reporting support study-specific quantitative comparison rather than a common pooled estimate (Hicks et al., 2019; Hu et al., 2024).

3.8 EVIDENCE-READINESS SYNTHESIS

Figure 5 compares five reporting dimensions in the selected investigations: independence of model validation, specification of the phenotypic reference, calibration or uncertainty information, reproducibility resources and operational evaluation. The chart provides a descriptive view of these dimensions and complements the organism-specific findings in Tables 1–4. Formal assessments of model bias and applicability require the full PROBAST+AI signalling questions and source-level evidence for each model evaluation (Moons et al., 2025).

Figure 5. Evidence-readiness matrix across verified studies

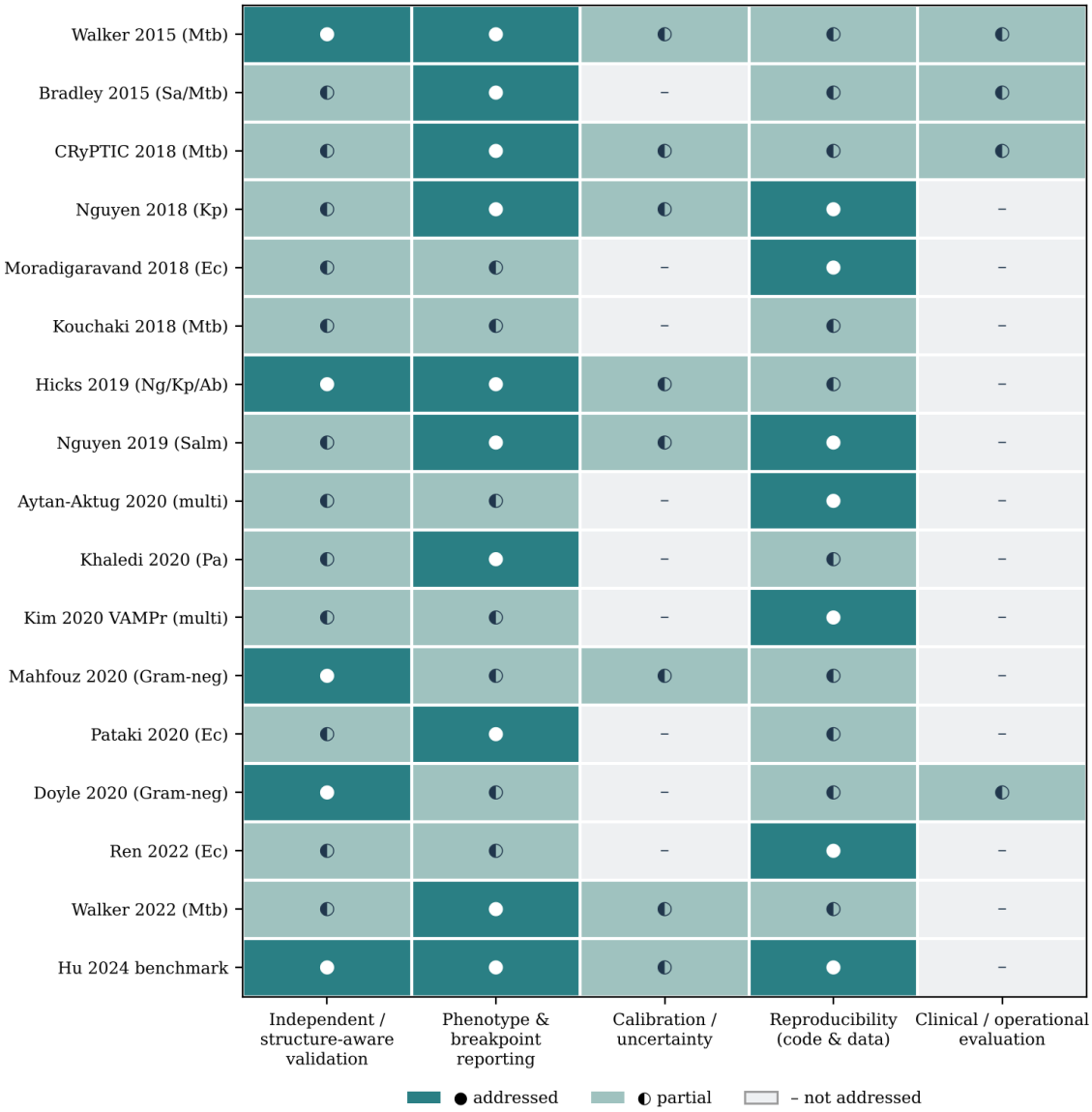


Figure 5. Evidence-readiness dimensions in selected genomic resistance prediction studies. Symbols denote the presence, partial description or absence of the specified reporting elements in the study summaries. The matrix is descriptive and distinct from formal PROBAST+AI risk-of-bias ratings.

3.9 ORGANISM AND ANTIMICROBIAL-CLASS INTERPRETATION

Mycobacterium tuberculosis. Clonal population structure and comparatively well-characterised chromosomal resistance mechanisms make curated mutation catalogues particularly useful for several drugs. Nonetheless, phenotypic discordance, resistance mutations of uncertain significance, and between-lineage variability require drug-specific interpretation. Evidence from international collections establishes substantial genotype–phenotype concordance for common first-line drugs but does not imply identical portability for newer or less frequently resistant agents (Bradley et al., 2015; CRyPTIC Consortium, 2022; CRyPTIC Consortium and the 100,000 Genomes Project, 2018; Kouchaki et al., 2019; Walker et al., 2015; Walker et al., 2022; World Health Organization, 2023). A complete external-validation synthesis should distinguish rifampicin and isoniazid from fluoroquinolones and second-line agents, chart the relevant

mutation catalogue version, and calculate sensitivity among resistant isolates with an explicit no-call fraction.

Enterobacterales. In Escherichia coli and Klebsiella pneumoniae, the pan-genome, mobile resistance determinants and changing plasmid backgrounds make generalisation across sites and time more challenging. Moradigaravand et al. evaluated multiple genomically derived predictors in E. coli, while Nguyen et al. predicted MICs across multiple agents in K. pneumoniae (Moradigaravand et al., 2018; Nguyen et al., 2018). Their results should not be pooled simply because the organisms share a family: drug identity, phenotype prevalence, feature representation and susceptibility method differ. For beta-lactams, fluoroquinolones, aminoglycosides and trimethoprim–sulfamethoxazole, evaluation must report each agent and organism separately, including new mobile-gene combinations and minority resistant classes.

Non-fermenters. Pseudomonas aeruginosa combines acquired determinants with chromosomal regulation, efflux, porin changes and adaptive physiology. Genomic-only models may not represent expression-dependent mechanisms captured in some multi-omic work (Khaledi et al., 2020). For Acinetobacter baumannii, dataset and laboratory heterogeneity remain important considerations, but the available targeted evidence is too limited to quantify an independent-test penalty. Carbapenem, aminoglycoside and polymyxin predictions should be evaluated in distinct groups because both biological mechanisms and phenotypic reference standards vary (Doyle et al., 2020; Hicks et al., 2019; Khaledi et al., 2020).

Gram-positive organisms. Staphylococcus aureus illustrates a setting in which a small set of established determinants can support high-performing predictions for certain antibiotics, as shown in Mykrobe evaluations (Bradley et al., 2015). This must not be extrapolated to all Gram-positive pathogens or all agents: lineage composition, inducible resistance, emerging mechanisms and clinical setting modify the applicability of a rule-based call. Enterococcal data in broad benchmarks provide complementary observations but should not be merged with S. aureus without organism-specific denominators (Bradley et al., 2015; Hu et al., 2024).

Other organisms and drug-specific tests. Studies of nontyphoidal Salmonella MIC prediction and Neisseria gonorrhoeae cross-dataset performance are particularly instructive. For Salmonella, MIC agreement and false-susceptibility rates should be assessed by agent and collection period, particularly when resistance is rare (Nguyen et al., 2019). In gonococci, ciprofloxacin and azithromycin involve different resistance genetic architectures and should not be merged into one average. Cross-collection and near-breakpoint behaviour may matter more than single split accuracy (Hicks et al., 2019; Pataki et al., 2020).

Across all subgroups, an internal-to-independent change should be presented only as a paired difference within a frozen model and identical task. For example, Δsensitivity = sensitivity_external − sensitivity_internal and Δspecificity = specificity_external − specificity_internal, each with denominators and uncertainty intervals. Where only an AUC, macro F1, essential agreement or aggregate percentage is available, the statistic should remain in its original column rather than being interpreted as a common sensitivity difference.

4. Discussion

4.1 PRINCIPAL FINDINGS

Three observations emerge from the selected studies and the multi-method benchmark. First, the performance of genomic resistance prediction depends heavily on how it is evaluated. The same models that reach a macro F1 of at least 0.9 in two-thirds of experiments under random partitions reach it in only a quarter under population-structure-aware partitions (Hu et al., 2024), and individual studies that held out independent datasets, laboratories or time periods showed the same direction of change (Doyle et al., 2020; Hicks et al., 2019; Nguyen et al., 2019; Ren et al., 2022). Strong internal performance therefore does not establish usefulness elsewhere, and the headline accuracies that dominate the field describe interpolation within a data distribution rather than transfer to new populations. Second, the size of this effect is organism-dependent: genomically clonal pathogens with well-characterised mechanisms, such as M. tuberculosis, transfer well, while genomically plastic Gram-negative pathogens,

such as P. aeruginosa and K. pneumoniae, transfer poorly (Hu et al., 2024). Third, available public genomes are geographically concentrated in high-income countries, with African and other low- and middle-income settings almost absent from model development and from the public genome collection as a whole (David et al., 2025), even though these settings carry the highest burden of resistance (Antimicrobial Resistance Collaborators, 2022).

Alongside these, the evidence relevant to clinical readiness is thin in a specific and consequential way. Discrimination is well studied; calibration, uncertainty and prospective evaluation are not (Kim et al., 2022). Where resistance is rare in the training data, false-susceptibility errors rise for exactly the determinants a clinician most needs to detect (Nguyen et al., 2019), and when the same data are analysed by different laboratories the predictions themselves disagree more than is often appreciated (Doyle et al., 2020). These are not reasons to abandon genomic prediction, which already has real value in defined settings such as tuberculosis, where catalogue-based prediction from sequence is used to guide regimen selection (CRyPTIC Consortium and the 100,000 Genomes Project, 2018; Walker et al., 2015; Walker et al., 2022). They are reasons to match the strength of a deployment claim to the strength of the evaluation that supports it.

4.2 RELATION TO PREVIOUS REVIEWS

Previous reviews have surveyed the algorithms used for genomic resistance prediction and have discussed clinical implementation in general terms (Kim et al., 2022). The EUCAST subcommittee set out the conditions that sequence-based susceptibility inference would need to meet before it could guide clinical decisions, and judged the evidence insufficient for most species at the time of writing (Ellington et al., 2017). The present review is complementary and more specific: rather than cataloguing methods, it compares internal against independent performance directly, charts the geography of the underlying isolates, and assesses readiness criterion by criterion. The benchmark that anchors much of the quantitative synthesis (Hu et al., 2024) provides, for the first time at scale, a like-for-like comparison of random and structure-aware evaluation across many species and drugs, and this review places that benchmark in the context of the single-organism and clinical evidence around it.

4.3 WHY THE GENERALISATION GAP ARISES

The mechanism behind the generalisation gap is well understood and worth stating plainly, because it determines the remedy. Bacterial genomes are inherited clonally over the short term, so isolates that share recent ancestry share most of their genome, including variation that has nothing to do with resistance. When resistance is also clustered by ancestry, which it usually is, a model can predict resistance accurately by recognising ancestry, without ever learning the causal determinant. A random split leaves relatives on both sides and rewards this shortcut; a structure-aware split removes the relatives and exposes it (Hicks et al., 2019; Hu et al., 2024). The same logic applies geographically and temporally: different regions and different periods carry different lineages and different determinants, so a model that has learned the lineages of one setting has not necessarily learned the biology that would let it transfer to another. Catalogue methods are partially protected from this particular failure, because they encode mechanism directly rather than learning it from correlation, and the benchmark found that the rule-based method was more consistent across evaluation designs, though it was limited by the completeness of the catalogue and could not be assessed where no catalogue entry existed (Hu et al., 2024). This suggests that mechanism-aware features and interpretable models, which tie a prediction to a causal determinant, are likely to transfer better than opaque correlational ones, a direction supported by the interpretable-model studies in this review (Kim et al., 2020; Pataki et al., 2020).

4.4 IMPLICATIONS FOR PRACTICE AND FOR EVALUATION

For laboratories and surveillance programmes considering genomic prediction, the practical implication is that the evaluation evidence must match the intended use. A model intended to classify isolates within a well-characterised, clonal pathogen in a setting resembling its training data may be adequately supported by careful internal evaluation. A model intended to be applied across hospitals, regions or time, or to a genomically plastic pathogen, requires

independent, structure-aware and ideally prospective evaluation before its outputs are trusted, and its calibration and false-susceptibility error must be known in the target setting. For developers, the implication is that the design of the evaluation is not a technical afterthought but the core scientific claim of the study, and that reporting it fully, including the handling of population structure, the phenotype reference standard and the error asymmetry, is what allows a reader to judge where the model can be used. The framework in Figure 2 and the checklist in Table 5 are intended to make these expectations concrete and reusable.

5. Evidence gaps and recommendations

The selected evidence and methodological literature identify several recurring evaluation weaknesses, each with a direct consequence for interpretation and a corresponding reporting or design remedy. Table 5 sets these out as a checklist for future studies and for reviewers and editors assessing them. The recommendations are deliberately framed so that they can be adopted without new methodology; most require only that existing good practice be applied and reported.

Table 5. Evidence gaps and recommended reporting for genomic AMR prediction studies.

Observed weakness	**Consequence for interpretation**	**Recommended reporting / design**
Random-split evaluation presented as evidence of generalisation	Population-structure leakage inflates apparent performance	Report population-structure-aware (lineage/phylogeny) evaluation; state how relatedness between train and test was controlled
Internal and independent performance not distinguished	Readers cannot tell interpolation from transfer	Report both; present paired internal and independent estimates for the same model and outcome where possible
Isolate origin unreported or conflated with author affiliation	Geographic applicability cannot be judged	Report the country and setting of isolate collection separately from author affiliation; flag unreported provenance
Under-representation of low- and middle-income settings	Models may not transfer to the highest-burden populations	State the geographic scope of applicability; prioritise inclusion of under-represented settings; share sequence-and-phenotype data
Phenotype reference standard and breakpoint version not stated	Predictions cannot be compared or reproduced	Report the AST method, the breakpoint standard and version, and the handling of intermediate categories
Calibration and uncertainty not reported	Clinicians cannot weigh a prediction against its confidence	Report calibration; provide probabilistic outputs and, where appropriate, an abstention mechanism for uncertain cases
Error asymmetry not reported with denominators	The clinically dangerous false-susceptibility rate is hidden	Report very major and major error rates with explicit denominators, by organism and drug
Reuse of the same public genomes across studies	Apparently separate findings are not independent	Identify datasets by accession; state overlap with previously used collections
Reproducibility not supported	Findings cannot be re-derived or extended	Share code, trained models, accessions and the evaluation split definitions

Absence of prospective/operational evaluation	Readiness for clinical use is unestablished	Where clinical use is claimed, evaluate predictions issued before the phenotype and audit them against the subsequent result

A concrete research agenda follows from these points. The highest priority is the generation and sharing of paired sequence-and-phenotype datasets from under-represented regions, without which the geographic validity gap cannot be closed (David et al., 2025). The second is the routine adoption of population-structure-aware evaluation as the default, with random-split results reported only as a secondary, clearly labelled comparison (Hicks et al., 2019; Hu et al., 2024). The third is the reporting of calibration, uncertainty and error asymmetry as standard outcomes rather than optional extras (Kim et al., 2022; Mahfouz et al., 2020; Nguyen et al., 2019). The fourth is a small number of genuinely prospective studies in defined clinical pathways, which are the only way to establish operational readiness (Ellington et al., 2017; Kim et al., 2022). None of these requires a methodological breakthrough; each requires that existing good practice be applied and reported consistently.

6. LIMITATIONS

The synthesis draws on selected published evaluations and major benchmarks, with emphasis on validation design and clinical interpretation. Variation in sampling frames, susceptibility methods, breakpoint versions and resistance prevalence influences comparisons between studies. Several investigations share public genome collections, and some published summary metrics are not accompanied by full organism–drug-level denominators. Geographic comparisons similarly depend on the availability of isolate collection metadata. These characteristics define the scope within which the findings can be interpreted and highlight the value of future comprehensive evidence extraction and independent testing across diverse bacterial populations.

7. CONCLUSIONS

Genomic antimicrobial resistance prediction provides useful information when the genetic determinants, phenotypic definitions and target populations are appropriately characterised. Evaluation design substantially influences reported accuracy, as demonstrated by the difference between random and phylogeny-aware testing within a large multi-species benchmark (Hu et al., 2024). Across pathogen groups, the most informative assessment combines genomic population structure, organism–drug-specific reference standards, geographic provenance and clinically meaningful error denominators. The integration of independent validation, uncertainty-aware reporting, reproducible bioinformatics pipelines and evaluation in intended clinical settings provides a practical basis for assessing the reliability and utility of genomic susceptibility predictions.

FUNDING

The authors gratefully acknowledge the Tertiary Education Trust Fund (TETFund) for supporting this research through the Institution-Based Research (IBR) intervention administered by Kashim Ibrahim University.

REFERENCES

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., & Wlodarski, M. A. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research, 51(D1), D690–D699. https://doi.org/10.1093/nar/gkac920

Antimicrobial Resistance Collaborators. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet, 399(10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0

Aun, E., Brauer, A., Kisand, V., Tenson, T., & Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. PLOS Computational Biology, 14(10), e1006434. https://doi.org/10.1371/journal.pcbi.1006434

Aytan-Aktug, D., Clausen, P. T. L. C., Bortolaia, V., Aarestrup, F. M., & Lund, O. (2020). Prediction of acquired antimicrobial

resistance for multiple bacterial species using neural networks. mSystems, 5(1), e00774-19. https://doi.org/10.1128/mSystems.00774-19

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., & Huang, B. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nature Communications, 6, 10063. https://doi.org/10.1038/ncomms10063

Clinical and Laboratory Standards Institute. (2026). Performance standards for antimicrobial susceptibility testing (36th ed.; CLSI supplement M100).

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., & Van Calster, B. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

CRyPTIC Consortium. (2022). A data compendium associating the genomes of 12,289 Mycobacterium tuberculosis isolates with quantitative resistance phenotypes to 13 antibiotics. PLOS Biology, 20(8), e3001721. https://doi.org/10.1371/journal.pbio.3001721

CRyPTIC Consortium and the 100,000 Genomes Project. (2018). Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. The New England Journal of Medicine, 379(15), 1403–1415. https://doi.org/10.1056/NEJMoa1800474

David, S., Caballero, J. D., Couto, N., Abudahab, K., Alikhan, N. F., & Yeats, C. (2025). Monitoring antimicrobial resistance trends from global genomics data: amr.watch. PLOS Global Public Health, 5(11), e0005256. https://doi.org/10.1371/journal.pgph.0005256

Doyle, R. M., O'Sullivan, D. M., Aller, S. D., Bruchmann, S., Clark, T., & Coello Pelegrin, A. (2020). Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. Microbial Genomics, 6(2), e000335. https://doi.org/10.1099/mgen.0.000335

Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., & Loo, V. G. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics, 17, 754. https://doi.org/10.1186/s12864-016-2889-6

Ellington, M. J., Ekelund, O., Aarestrup, F. M., Canton, R., Doumith, M., & Giske, C. (2017). The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clinical Microbiology and Infection, 23(1), 2–22. https://doi.org/10.1016/j.cmi.2016.11.012

European Committee on Antimicrobial Susceptibility Testing. (2026). Breakpoint tables for interpretation of MICs and zone diameters (Version 16.1).

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., & Tolstoy, I. (2019). Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype–phenotype correlations in a collection of isolates. Antimicrobial Agents and Chemotherapy, 63(11), e00483-19. https://doi.org/10.1128/AAC.00483-19

Hicks, A. L., Wheeler, N., Sánchez-Busó, L., Rakeman, J. L., Harris, S. R., & Grad, Y. H. (2019). Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. PLOS Computational Biology, 15(9), e1007349. https://doi.org/10.1371/journal.pcbi.1007349

Hu, K., Meyer, F., Deng, Z. L., Asgari, E., Kuo, T. H., Münch, P. C., & McHardy, A. C. (2024). Assessing computational predictions of antimicrobial resistance phenotypes from microbial genomes. Briefings in Bioinformatics, 25(3), bbae206. https://doi.org/10.1093/bib/bbae206

Khaledi, A., Weimann, A., Schniederjans, M., Asgari, E., Kuo, T. H., & Oliver, A. (2020). Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine

learning-enabled molecular diagnostics. EMBO Molecular Medicine, 12(3), e10264. https://doi.org/10.15252/emmm.201910264

Kim, J., Greenberg, D. E., Pifer, R., Jiang, S., Xiao, G., & Shelburne, S. A. (2020). VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. PLOS Computational Biology, 16(1), e1007511. https://doi.org/10.1371/journal.pcbi.1007511

Kim, J. I., Maguire, F., Tsang, K. K., Gouliouris, T., Peacock, S. J., & McAllister, T. A. (2022). Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. Clinical Microbiology Reviews, 35(3), e00179-21. https://doi.org/10.1128/cmr.00179-21

Kouchaki, S., Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., & Peto, T. E. A. (2019). Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics, 35(13), 2276–2282. https://doi.org/10.1093/bioinformatics/bty949

Mahfouz, N., Ferreira, I., Beisken, S., von Haeseler, A., & Posch, A. E. (2020). Large-scale assessment of antimicrobial resistance marker databases for genetic phenotype prediction: a systematic review. Journal of Antimicrobial Chemotherapy, 75(11), 3099–3108. https://doi.org/10.1093/jac/dkaa257

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., & Dhiman, P. (2025). PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505. https://doi.org/10.1136/bmj-2024-082505

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., & Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLOS Computational Biology, 14(12), e1006258. https://doi.org/10.1371/journal.pcbi.1006258

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., & Olson, R. (2018). Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Scientific Reports, 8(1), 421. https://doi.org/10.1038/s41598-017-18972-w

Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., & Stevens, R. L. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella. Journal of Clinical Microbiology, 57(2), e01260-18. https://doi.org/10.1128/JCM.01260-18

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., & Mulrow, C. D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Pataki, B. Á., Matamoros, S., van der Putten, B. C. L., Remondini, D., Giampieri, E., & Aytan-Aktug, D. (2020). Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. Scientific Reports, 10(1), 15026. https://doi.org/10.1038/s41598-020-71693-5

Ren, Y., Chakraborty, T., Doijad, S., Falgenhauer, L., Falgenhauer, J., & Goesmann, A. (2022). Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics, 38(2), 325–334. https://doi.org/10.1093/bioinformatics/btab681

Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., & Bradley, P. (2015). Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. The Lancet Infectious Diseases, 15(10), 1193–1202. https://doi.org/10.1016/S1473-3099(15)00062-6

Walker, T. M., Miotto, P., Köser, C. U., Fowler, P. W., Knaggs, J., & Iqbal, Z. (2022). The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: a genotypic analysis. The Lancet Microbe, 3(4), e265–e273.

https://doi.org/10.1016/S2666-5247(21)00301-3

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., & Reitsma, J. B. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

World Health Organization. (2023). Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance (2nd ed.).

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., & Lund, O. (2012). Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy, 67(11), 2640–2644. https://doi.org/10.1093/jac/dks261