e-ISSN: 0976-822X, p-ISSN:2961-6042

Available online on http://www.ijcpr.com/

International Journal of Current Pharmaceutical Review and Research 2026; 18(9); 769-784

Original Research Article

Genomic Antimicrobial Resistance Architecture of Orthopaedic-Associated *Staphylococcus Aureus* in NCBI Pathogen Detection: A Project-Aware Secondary Analysis

Khandayatray P.[1], Brahma N.[2]

[1]Department of Microbiology, PGIMER & Capital Hospital, Bhubaneswar, Odisha, India
[2]Department of Orthopedics, PGIMER & Capital Hospital, Bhubaneswar, Odisha, India

Received: 01-06-2026 / Revised: 15-07-2026 / Accepted: 21-08-2026
Corresponding author: Dr. Priyadarsani Khandayatray
Conflict of interest: Nil

Abstract

Objective: To characterise the genomic antimicrobial-resistance (AMR) architecture of orthopaedic-associated *Staphylococcus aureus* in NCBI Pathogen Detection and determine whether apparent source- or comparator-associated differences persist after accounting for sequencing-project structure and lineage-correlated genomic background.

Methods: We performed a secondary analysis of frozen NCBI Pathogen Detection release PDG000000073.1274. The primary cohort comprised 513 high-confidence orthopaedic-associated *Staphylococcus aureus* isolates, with 10,644 strictly defined non-orthopaedic clinical comparators. Seven prespecified genomic AMR classes were evaluated using descriptive prevalence and co-occurrence analyses, temporal and geographic models, orthopaedic source-group comparisons, BioProject-aware regression, within-BioProject conditional logistic regression, and an accessory-genome principal-component lineage-proxy sensitivity analysis.

Results: Macrolide-associated determinants were present in 40.0% of orthopaedic isolates and methicillin-associated determinants in 38.0%; 43.5% carried none of the seven primary AMR classes. The unstratified source-group PERMANOVA was significant ($p=0.0005$; $R^2=0.0467$), but dispersion differed markedly and the BioProject-stratified sensitivity was null ($p=0.671$). In strict within-BioProject comparison of 1,414 isolates across 25 shared BioProjects, no primary AMR class remained significant after Holm correction. A post hoc lineage-correlated accessory-genome sensitivity analysis did not alter this corrected conclusion. No robust nonlinear temporal trend was detected for the five common AMR classes examined.

Conclusion: Orthopaedic-associated *Staphylococcus aureus* in this public genomic repository showed substantial but heterogeneous AMR architecture. However, apparent source- and comparator-associated differences weakened after stringent control for BioProject and lineage-correlated genomic background, indicating that repository composition can materially influence pooled genomic AMR comparisons.

Keywords: Orthopaedic Infection; *Staphylococcus aureus*; Antimicrobial Resistance; Genomic Epidemiology; NCBI Pathogen Detection; BioProject; Population Structure.

DOI: 10.25258/ijcpr.18.9.110

This is an Open Access article that uses a funding model which does not charge readers or their institutions for access and distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0) and the Budapest Open Access Initiative (http://www.budapestopenaccessinitiative.org/read), which permit unrestricted use, distribution, and reproduction in any medium, provided original work is properly credited.

Introduction

Musculoskeletal and orthopaedic infections caused by *Staphylococcus aureus* span osteomyelitis, septic arthritis, fracture-related infection, prosthetic-joint infection and other device-associated syndromes, and remain clinically consequential because infection control may require prolonged, multidisciplinary management. A recent systematic review and meta-analysis identified *S. aureus* as a leading pathogen across musculoskeletal infections and documented substantial variation in antimicrobial resistance between settings. [1] In the multinational AO Trauma bone-infection registry, culture-confirmed *S. aureus* infection included fracture-related infection, prosthetic-joint infection and osteomyelitis; fewer than two-thirds of participants were classified as cured at 12 months, underscoring the persistent burden of these infections despite treatment.2 Genomic studies have further shown that clinical outcome cannot be explained by a single bacterial feature across the heterogeneous spectrum of bone and joint infection. [3] Antimicrobial resistance adds another layer of complexity. Methicillin resistance remains

clinically important, while resistance to macrolides, lincosamides, aminoglycosides, fluoroquinolones and other non-beta-lactam agents can constrain therapeutic options and complicate surveillance. However, resistance estimates in musculoskeletal infection are not uniform: pooled and local studies demonstrate marked heterogeneity by geography, infection type, period and sampling frame. [1,2] This variation makes a single universal resistance percentage potentially misleading and highlights the need to distinguish local phenotypic susceptibility patterns from broader genomic resistance architecture. The distinction is especially important when genomic determinants are analysed, because the presence of a resistance gene or mutation represents a genomic prediction or mechanism rather than direct measurement of antimicrobial susceptibility in the corresponding clinical isolate.

Whole-genome sequencing has expanded the study of *S. aureus* beyond conventional susceptibility testing by allowing resistance determinants, sequence types, clonal background and broader accessory-genome features to be examined together. A systematic review of *S. aureus* genomics in bone and joint infection found substantial methodological and clinical heterogeneity, with no consistent adult genomic marker explaining adverse outcomes across studies. [3] More recent orthopaedic sequencing studies similarly demonstrate diverse sequence types and resistance-gene profiles rather than a single orthopaedic lineage. [4] In prosthetic-joint infection, whole-genome comparisons have not identified a PJI-specific phylogenetic cluster, [5] while lineage-aware genome-wide analyses of orthopaedic implant infections have shown that apparent group-associated variants can diminish after population structure is considered. [6] These findings establish the value of genomics, but also show that clinical-source associations require careful control for bacterial background.

Public genomic repositories offer an opportunity to extend such analyses beyond single institutions. NCBI Pathogen Detection is a centralized system that integrates pathogen genomic sequences and contextual metadata from numerous surveillance and research efforts and applies automated analyses, including AMRFinderPlus-based identification of antimicrobial-resistance, stress-response and virulence determinants. [7] The curated AMRFinderPlus databases provide a structured framework for detecting known resistance genes and point mutations across large genome collections. [8] This scale enables secondary genomic epidemiology across many submitting projects and locations. At the same time, NCBI Pathogen Detection is dependent on data submitted by participating public-health agencies and researchers; consequently, the resulting isolate collection should be interpreted as a heterogeneous public repository rather than as a probability sample or a population-representative clinical antibiogram. [7] Large public-database studies have already used NCBI Pathogen Detection to describe *S. aureus* resistance-gene patterns across broad human, animal, environmental and geographic sources. [9,10] These studies demonstrate the analytical value of the repository, but they also make clear that source and geographic distributions are uneven. More generally, bacterial genomic analyses are vulnerable to confounding when sampling is correlated with phylogeny or other population structure. In a multi-pathogen analysis, biased sampling driven by bacterial population structure materially distorted antimicrobial-resistance prediction and was not rescued simply by increasing sample size. [11] Population-structure-aware association methods were developed precisely because lineage can account for substantial phenotypic variation in bacterial datasets. [12] Surveillance methodology also shows that systematic processes determining which infections are detected, sequenced and characterized can bias inferred frequencies. [13] Thus, for orthopaedic-source comparisons in a public repository, apparent AMR differences may reflect not only clinical niche but also the composition of the sequencing projects, locations and lineages contributing the isolates.

Against this background, the present study was designed to characterise the prevalence, co-occurrence and distribution of major genomic antimicrobial-resistance determinants among high-confidence orthopaedic-associated *S. aureus* isolates in NCBI Pathogen Detection; evaluate temporal, geographic and orthopaedic specimen-source variation; compare orthopaedic-associated isolates with a strictly defined non-orthopaedic clinical comparator; and determine whether apparent associations remain robust under BioProject-aware and lineage-correlated population-structure adjustment. The aim was therefore not to presume a unique orthopaedic resistance signature, but to distinguish descriptive genomic AMR architecture from associations that may arise from the structure of a heterogeneous public repository.

Materials and Methods

Study design and reporting framework: This study was an observational secondary genomic-epidemiology analysis of publicly deposited *Staphylococcus aureus* isolate data. The analytical unit was the microbial isolate; no assumption was made that one isolate represented one individual patient. Core cohort construction, specimen-source classification, exclusion rules, genomic

antimicrobial-resistance (AMR) class definitions and the principal comparator/sensitivity analyses were specified and frozen before interpretation of AMR outcomes. The later accessory-genome lineage-proxy analysis was added as a post hoc robustness analysis after direct typing fields proved insufficient; it did not redefine cohort membership, AMR classes or primary outcomes. No a priori sample-size calculation was performed because all eligible records in the frozen repository snapshot were included. The reporting framework was informed by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement and the STROME-ID extension for infectious-disease molecular epidemiology. [15,16]

Data source and frozen extraction snapshot: The primary source was the National Center for Biotechnology Information (NCBI) Pathogen Detection system, which integrates bacterial genomic sequences and contextual metadata from multiple surveillance and research submissions and reports AMR, stress-response and virulence annotations generated by AMRFinderPlus. A fixed *S. aureus* organism-group snapshot was used to prevent changes in the continuously updated repository from altering the analysis after cohort construction. The extraction was performed on 4 September 2026 using Pathogen Detection release PDG000000073.1274 (release directory dated 3 September 2026). The source metadata file was PDG000000073.1274.metadata.tsv, a tab-delimited UTF-8 file containing 171,370 records and 67 fields excluding the header. Submitted antimicrobial-susceptibility testing (AST) data were obtained separately from asts (1).csv. The original files were retained unchanged, and file sizes and SHA-256 checksums were recorded in the data-acquisition manifest. [7,14]

The master isolate dataset retained, where available, the Pathogen Detection target accession, BioSample, assembly and SRA accessions, BioProject, isolate identifiers, taxonomy, source type, host, epidemiological type, host disease, isolation source, collection date, geographic location, AMR genotypes, AMR core genotypes, AMRFinderPlus analysis type and version, Reference Gene Catalog version, and virulence and stress genotypes. Raw variables were preserved separately from normalized variables so that every analytical record could be traced back to the original repository row.

Organism, human-source and clinical eligibility: Records were first restricted to *S. aureus* using NCBI species TaxID 1280. Human-source evidence was then assigned hierarchically. H1 denoted records explicitly identified as Human in the source_type field; H2 denoted records in which the host field unambiguously identified Homo sapiens or an equivalent human term. H3 denoted records labelled clinical but lacking definitive human host evidence, H0 denoted clearly non-human sources, and HU denoted insufficiently classifiable human-source status. The primary human clinical pool required H1 or H2 evidence together with epi_type classified as clinical. H3, H0 and HU records were not admitted to the primary human cohort.

Isolation-source text was preserved verbatim and also normalized reproducibly by case folding, whitespace collapse and punctuation/hyphen standardization. Clinically meaningful words were not removed. Collection source, human-source evidence and clinical status were maintained as separate variables; a human source alone was not assumed to represent active infection.

Orthopaedic specimen-source classification: A prespecified rule-based, high-specificity classifier was applied to normalized isolation-source text, with host-disease context used only where defined in the frozen rule ledger. The primary orthopaedic cohort (ORTHO_STRICT) comprised six mutually coded source groups: O1, joint/synovial; O2, bone/osteomyelitis; O3, prosthesis/PJI-associated source; O4, orthopaedic fixation/device; O5, spinal musculoskeletal; and O6, other high-confidence musculoskeletal source. Strict inclusion required explicit source terminology such as synovial or joint fluid/tissue, bone biopsy/tissue or osteomyelitis, prosthetic-joint/periprosthetic wording, explicit orthopaedic hardware/fixation wording, or vertebral/spinal musculoskeletal wording. The O3 label described the submitted specimen source and was not treated as proof that formal diagnostic criteria for prosthetic joint infection had been met.

Context-protection and exclusion overrides were applied before inclusion. Cerebrospinal fluid or spinal fluid was classified as a non-orthopaedic central-nervous-system source rather than as spinal musculoskeletal infection. Bone marrow, craniofacial bone terms such as mandible or cheekbone, and clearly non-orthopaedic devices (including prosthetic cardiac valves, vascular grafts, dental or breast implants, catheters and pacemakers) were prevented from entering the strict orthopaedic cohort. Explicit screening, surveillance, carriage, nasal/nares sampling and colonisation terminology was kept outside the infected clinical cohorts.

A broader sensitivity cohort (ORTHO_EXPANDED) retained ORTHO_STRICT and added only moderate-specificity, context-supported records coded OA, including selected implant-sonicate/hardware descriptions or anatomically specific aspirates with orthopaedic context. Generic wound, tissue, aspirate or deep-tissue wording without explicit

orthopaedic context was not promoted into the orthopaedic cohort. Classifier implementation underwent internal dual-pass semantic quality control. A stratified 1,250-record set spanning strict orthopaedic, comparator, ambiguous/misleading and other-source strata was prepared for independent external review. Independent clinician adjudication had not been completed at this manuscript stage. Accordingly, no classifier sensitivity, specificity, predictive value or kappa estimate was presented as externally validated clinical performance; the classifier is described only as a rule-based, high-specificity specimen-source classification.

Strict non-orthopaedic clinical comparator: The primary comparator (COMPARATOR_STRICT) was restricted to high-confidence human clinical *S. aureus* records with clear non-orthopaedic sources. Prespecified eligible categories comprised respiratory, urinary, central-nervous-system, intra-abdominal and clearly non-musculoskeletal skin/soft-tissue sources. Blood, bacteraemia and bloodstream isolates were excluded because bacteraemia may occur secondary to bone or joint infection and could therefore contaminate the non-orthopaedic comparator. Generic wound, tissue, deep-tissue, aspirate and unspecified clinical specimens were also excluded when their anatomical origin could not be established. Surveillance/colonisation specimens, ambiguous musculoskeletal sources and non-orthopaedic device sources were not included in COMPARATOR_STRICT.

Duplicate assessment and analytical independence: Potential duplication and non-independence were evaluated using Pathogen Detection target accession, BioSample, assembly accession, SRA Run, isolate identifiers, collection date/source and BioProject. Exact technical duplicates were to be resolved deterministically by retaining one canonical record. Repeated BioSample or potentially re-sequenced records that could still represent biologically distinct isolates were flagged rather than automatically discarded when individual-level linkage was unavailable. The final analytical unit therefore remained the deposited microbial isolate, with residual dependence addressed principally through BioProject-aware models and project-restricted sensitivity analyses.

Genomic AMR annotation and outcome definitions: AMR outcomes were derived from the NCBI Pathogen Detection AMRFinderPlus annotations. The original AMR_genotypes and AMR_genotypes_core fields, AMRFinderPlus analysis type, AMRFinderPlus version and Reference Gene Catalog version were retained. The main class indicators were generated from AMR_genotypes_core using a frozen determinant-to-class dictionary. Only high-confidence COMPLETE calls and resistance-associated POINT calls were accepted for the seven primary class indicators; partial, HMM-only, mistranslation and other non-primary-quality calls were not used to create these indicators. [14]

Seven primary genomic AMR classes were prespecified. Methicillin markers comprised mecA or mecC. Macrolide markers comprised erm(A), erm(B), erm(C), erm(T), msr(A) and mph(C). Lincosamide markers comprised erm(A/B/C/T), lnu(A), lnu(B), lsa(E) and vga(A). Aminoglycoside markers comprised prespecified aac, aph and ant family determinants and aadD1. Tetracycline markers comprised tet(K), tet(L), tet(M) and tet(C). Quinolone markers were prespecified resistance-associated mutations in gyrA, parC or parE. Rifamycin markers were prespecified resistance-associated rpoB substitutions. A separate frozen dictionary mapped additional determinants to secondary classes, which were retained primarily for descriptive analysis when event counts were too small for stable modelling.

AMR missingness was handled explicitly. The literal value <NONE> was interpreted as an available AMRFinderPlus analysis in which no listed relevant AMR genotype was detected; an empty or unavailable AMR field was treated as missing output and was never recoded as a negative result. NUCLEOTIDE and COMBINED AMRFinderPlus analysis modes were preserved separately. A sensitivity analysis restricted the main analyses to COMBINED-mode records because the underlying annotation mode differs from nucleotide-only processing. [7,14]

Temporal, geographic and BioProject variables: Collection year was derived only from the submitted collection_date field when an interpretable four-digit year was present; NCBI target creation date was not substituted for missing collection time. Geographic location was normalized from geo_loc_name to specimen-collection country using a reproducible alias dictionary and then mapped to continent. Country was not inferred from the submitting laboratory, BioProject owner or author affiliation. BioProject accession was retained as the principal variable representing submission/research-project structure.

Descriptive and co-occurrence analyses: Binary AMR-class prevalence was summarized as n/N and percentage with 95% Wilson binomial confidence intervals. Genomic resistance-class burden was defined as the number of the seven primary AMR classes present in each isolate (range 0-7); this measure was not labelled phenotypic multidrug resistance. Common combinations were tabulated. Pairwise similarity between AMR classes was measured with the Jaccard coefficient using class-

by-isolate presence vectors, and average-linkage hierarchical clustering was used to order the class-similarity matrix. Co-occurrence visualizations were used only for combinations with interpretable frequency.

Orthopaedic source-group analyses: Class-specific distributions were evaluated principally across the three adequately represented source groups: bone/osteomyelitis, joint/synovial and prosthesis/PJI-associated source. Smaller fixation/device, spinal and other musculoskeletal groups were retained descriptively rather than forced into unstable independent models. Overall class-specific source heterogeneity was assessed with contingency-table chi-square tests, followed where appropriate by pairwise comparisons with multiplicity control.

For multivariate AMR-profile structure, each isolate in the three main source groups was represented by a seven-dimensional binary vector. Pairwise Jaccard distances were calculated and classical principal coordinates analysis (PCoA) was obtained by double-centering the squared distance matrix and eigendecomposition. Source-group separation was assessed with PERMANOVA using 1,999 permutations (random seed 20260904), with pseudo-F and R2 calculated from between- and within-group distance sums of squares. To examine project confounding, a sensitivity analysis permuted source labels only within BioProjects that contributed at least two orthopaedic source groups. Because PERMANOVA can be influenced by heterogeneous within-group dispersion, dispersion was separately assessed as distance to the group centroid in positive PCoA coordinate space (up to the first 10 positive axes) and compared by analysis of variance.

Resistance-class burden across the three principal source groups was modelled with Poisson generalized estimating equations (GEE). Bone/osteomyelitis was the reference group; joint/synovial and prosthesis/PJI-associated source were entered as indicators. Collection year was scaled per five years as (year - 2015)/5, continent was included with Europe as the reference when available, and BioProject was used as the clustering group with an independence working correlation structure and robust sandwich covariance. Incidence-rate ratios (IRRs) with 95% confidence intervals were derived, and a joint Wald test evaluated the source-group block. This specification accounted for within-project dependence but was not interpreted as eliminating all between-project confounding; source contrasts lacking adequate within-project overlap were therefore treated as exploratory.

Temporal and geographic models: Temporal associations for major AMR classes were modelled using collection year scaled per five-year increase, with continent adjustment and BioProject-clustered covariance. To test for non-linear temporal patterns without selecting eras on the basis of AMR outcomes, restricted cubic spline sensitivity models were limited to primary classes with prevalence of at least 20% and to the better represented 2000-2025 collection period selected from sample-density inspection. A four-degree-of-freedom restricted cubic spline basis was generated using patsy (cr(x, df=4)-1). Binomial generalized linear models were adjusted for continent and fitted with BioProject-clustered covariance. Overall evidence for temporal non-linearity was evaluated by a Wald test of the spline terms; standardized model predictions across the observed continent distribution were generated for visualization. Geographic distributions were first described by country and continent. Formal country-level comparisons were restricted to locations with approximately 30 or more orthopaedic isolates and adequate outcome counts; broader regional/continental comparisons were preferred when country-level data were sparse. Formal continent contrasts were fitted with binomial models adjusted for collection year and BioProject-clustered covariance, using Europe as the reference where represented. Inferential estimates were suppressed when an event or non-event cell contained fewer than five isolates. Benjamini-Hochberg false-discovery-rate adjustment was applied to the family of formal geographic comparisons. Because cluster-robust covariance addresses correlated observations but does not by itself remove between-project confounding when geography and BioProject are strongly aligned, these geographic models were treated as exploratory repository-level contrasts rather than estimates of regional or national resistance prevalence.

Orthopaedic-versus-comparator project-aware modelling: The comparator outcome was coded as orthopaedic-associated source = 1 and COMPARATOR_STRICT = 0. Initial pooled class-specific models examined each prespecified primary AMR class, adjusting for collection year and continent with BioProject-clustered covariance. This covariance specification accounted for within-project dependence but did not remove unmeasured between-project confounding. A joint pooled binomial model then entered all seven primary AMR classes simultaneously with the same covariates. Adjusted odds ratios (aORs) and 95% confidence intervals were calculated. Holm correction was applied across the seven AMR coefficients. Multicollinearity among the seven AMR predictors was evaluated with variance-inflation factors.

Project-composition sensitivity analyses first restricted the pooled comparator to BioProjects contributing at least one orthopaedic and at least one strict comparator isolate, thereby improving project overlap while retaining BioProject-clustered covariance. A stricter conditional analysis (Section 2.13) then identified effects solely from orthopaedic-versus-comparator contrasts within the same BioProjects and was treated as the strongest available control for project-level factors constant within project. Additional prespecified robustness analyses used the COMBINED AMRFinderPlus subset, ORTHO_EXPANDED, a modern-period subset, removal of the largest orthopaedic-contributing BioProject, and leave-one-shared-BioProject-out refitting for key associations. Robustness was evaluated using effect direction, effect magnitude, confidence intervals and project influence rather than retention of nominal statistical significance alone.

Within-BioProject conditional analysis: A stricter within-project analysis used conditional logistic regression (statsmodels ConditionalLogit). Only BioProjects containing both ORTHO_STRICT and COMPARATOR_STRICT records and isolates with usable collection year contributed to these models. BioProject was specified as the stratum, such that effect estimates were identified from orthopaedic-versus-comparator contrasts within the same contributing project rather than between projects. Collection year was scaled as (year - 2015)/5.

Seven single-class-plus-year models were fitted, followed by a joint model containing all seven primary AMR classes and year. Holm correction was applied across the seven AMR coefficients for each model family. This conditional specification was prespecified as the strongest control for project-level factors that were constant within BioProject.

Post hoc lineage-correlated population-structure sensitivity: Direct sequence type (ST) or clonal-complex information was not available for the full cohort because the computed_types field was unpopulated in the frozen *S. aureus* snapshot. No missing ST or clonal-complex values were inferred. To probe whether the strict within-BioProject conclusion was sensitive to lineage-correlated genomic background, a post hoc population-structure robustness analysis constructed an accessory-genome proxy from the non-AMR AMRFinderPlus virulence_genotypes and stress_genotypes fields. This analysis was explicitly sensitivity-oriented rather than a replacement for direct MLST, clonal-complex assignment or phylogeny. Presence/absence features occurring in 1%-99% of the combined orthopaedic/comparator cohort were retained. Principal-component analysis (PCA; scikit-learn, random_state=42) was applied to the binary feature matrix, with up to 15 components calculated. The primary background set was defined as the smallest number of PCs exceeding 60% cumulative explained variance; a 10-PC specification was retained as sensitivity analysis.

Lineage correlation of the accessory-genome proxy was evaluated only against ST labels explicitly recoverable from submitter-entered strain, isolate_identifiers or isolation_source fields. A regular-expression rule extracted a single explicit ST number and did not infer labels where none were stated. Evaluation was restricted to ST groups with at least five labelled isolates. The first 10 accessory PCs were assessed using a multigroup pseudo-F statistic and R2, with 4,999 label permutations (seed 42). Predictive separation of ST labels was additionally evaluated using logistic regression (maximum 3,000 iterations) and five-fold stratified cross-validation with shuffling (random_state=42), reporting balanced accuracy. These checks established that the PCs were strongly lineage-correlated in the limited labelled subset; they were not treated as direct or externally validated MLST/clonal-complex assignments.

Within shared BioProjects, a year-only conditional source model was compared with a year-plus-primary-PC model by likelihood-ratio chi-square testing. The principal lineage-proxy AMR models then fitted each AMR class separately with collection year and the primary accessory-PC set within BioProject strata, followed by a joint seven-class model with the same covariates. Holm correction was applied across the seven AMR-class coefficients. The full analysis was repeated using 10 PCs to assess sensitivity to the number of background components. A bone-versus-joint within-project population-structure analysis was treated as exploratory because project overlap was limited; a prosthesis-specific lineage model was not forced when within-project representation was insufficient.

AST linkage: Phenotypic AST data were processed as a separate long-format dataset in which one row represented one isolate-antimicrobial test result. Exact *S. aureus* AST records were linked to the isolate master dataset using Pathogen Detection isolate and/or BioSample identifiers. Submitted resistance phenotype, MIC or disk-diffusion measurement, testing standard and laboratory method were retained where available. Harmonization was restricted to unambiguous susceptible (S) and resistant (R) calls; intermediate or nonsusceptible categories were not silently recoded as resistant. Because paired orthopaedic AST coverage was too sparse for stable diagnostic-performance estimation, AST was retained only as a descriptive supplementary component, and no inferential sensitivity, specificity, positive

predictive value or negative predictive value was calculated.

Missing data, diagnostics and multiplicity: Analyses used analysis-specific complete cases; source classification, AMR-result availability, collection year and geographic location were not imputed. The eligible and analyzed denominator for each major model was retained in the statistical outputs. Model checks included convergence, complete or quasi-complete separation, sparse event cells, influential observations, cluster count, multicollinearity and extreme predicted probabilities. Models failing basic diagnostics were suppressed or downgraded rather than forced.

All inferential tests were two-sided. A nominal alpha of 0.05 was used, but interpretation prioritized effect sizes, 95% confidence intervals, multiplicity correction and cross-specification robustness. Holm adjustment was used for the prespecified seven-class AMR hypothesis families; Benjamini-Hochberg adjustment was used for the formal geographic-comparison family. Nominal associations that did not survive the relevant correction were not treated as confirmed findings.

Reproducibility and software: Cohort construction and statistical analyses were performed programmatically in Python 3.13.5. The final execution environment included NumPy 2.3.5, SciPy 1.17.0, statsmodels 0.14.6, patsy 1.0.2, pandas 2.2.3, scikit-learn 1.8.0 and Matplotlib 3.10.8. The extraction/classification pipeline used random seed 20260904; PERMANOVA used the same seed, and the lineage-proxy permutation/cross-validation procedures used seed 42. Frozen source files, SHA-256 hashes, cohort and exclusion logs, statistical scripts, model outputs, figures and result-lock manifests were retained. Key cohort counts and headline statistical outputs were independently recalculated through a second computational route before the result locks were accepted.

Ethics and data governance

This study used only publicly accessible, de-identified microbial genomic and contextual metadata deposited in NCBI Pathogen Detection. The investigators did not recruit participants, intervene in clinical care, collect new human specimens, access private patient records or attempt to re-identify individuals. An institution-specific statement regarding ethics review, exemption or waiver should be confirmed by the authors before submission; no ethics approval number or waiver is asserted in this draft.

Results

Cohort characteristics: The frozen NCBI Pathogen Detection snapshot contained 171,370 metadata records, of which 170,711 were taxonomically verified as Staphylococcus aureus. High-confidence human origin was assigned to 109,048 isolates, 108,850 met the human clinical definition, and 66,143 had source information sufficient for source classification. The primary ORTHO_STRICT cohort comprised 513 orthopaedic-associated isolates; ORTHO_EXPANDED contained 535 isolates and COMPARATOR_STRICT contained 10,644 clearly non-orthopaedic clinical isolates. Within ORTHO_STRICT, 289 isolates were bone/osteomyelitis-associated, 147 joint/synovial-associated, 59 prosthesis/PJI-associated by specimen source, 15 fixation/device-associated, two spinal, and one other high-confidence musculoskeletal isolate. Collection year was available for 504/513 isolates (1970-2025), 21 countries were represented, and the cohort originated from 103 BioProjects. AMRFinderPlus results were available for all 513 isolates; 408 were COMBINED analyses and 105 were NUCLEOTIDE analyses (Table 1; Figure 1).

Table 1: Cohort and metadata characteristics.

Characteristic	N	Note
ORTHO_STRICT	513	Primary high-specificity orthopaedic cohort
ORTHO_EXPANDED	535	Strict cohort plus adjudicated ambiguous orthopaedic-source records
COMPARATOR_STRICT	10,644	Clearly defined non-orthopaedic human clinical isolates
Bone/osteomyelitis-associated	289	ORTHO_STRICT source subgroup
Joint/synovial-associated	147	ORTHO_STRICT source subgroup
Prosthesis/PJI-associated source	59	Specimen-source label; not necessarily confirmed PJI
Fixation/device-associated	15	ORTHO_STRICT source subgroup
Spinal / other musculoskeletal	2 / 1	Descriptive only due small n
Collection year available	504/513	Range 1970-2025
Countries represented	21	Deposited-isolate geography
BioProjects represented	103	ORTHO_STRICT
AMRFinderPlus result available	513/513	COMBINED 408; NUCLEOTIDE 105

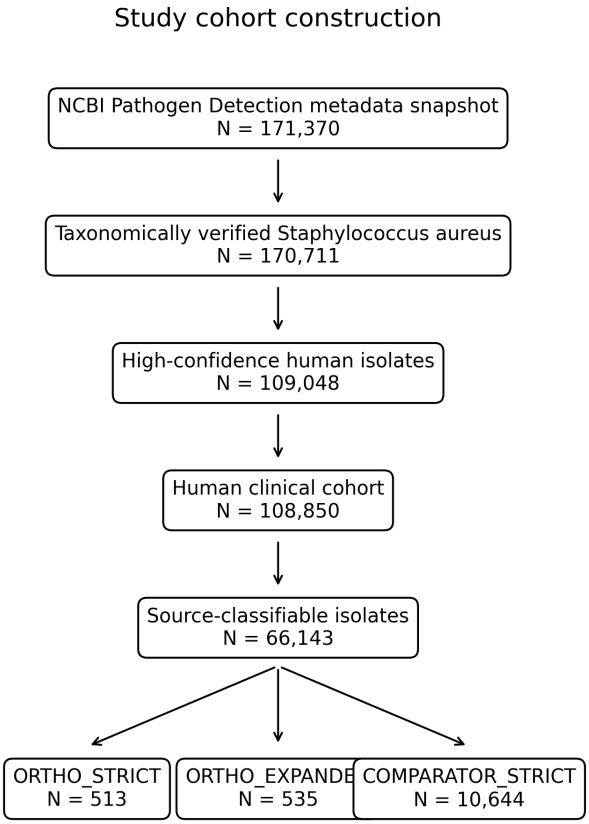


Figure 1: Study cohort construction from the frozen NCBI Pathogen Detection *Staphylococcus aureus* metadata snapshot. ORTHO_STRICT prioritised high-specificity orthopaedic specimen-source classification; ORTHO_EXPANDED additionally included adjudicated ambiguous orthopaedic-source records.

Genomic AMR prevalence and resistance architecture: Among the seven prespecified primary AMR classes, macrolide-associated determinants were identified in 205/513 isolates (40.0%, 95% CI 35.8-44.3), followed by methicillin-associated determinants in 195/513 (38.0%, 95% CI 33.9-42.3), lincosamide-associated determinants in 177/513 (34.5%, 95% CI 30.5-38.7), aminoglycoside-associated determinants in 167/513 (32.6%, 95% CI 28.6-36.7), and quinolone-associated determinants or curated resistance mutations in 166/513 (32.4%, 95% CI 28.5-36.5). Tetracycline-associated determinants occurred in 49/513 isolates (9.6%, 95% CI 7.3-12.4) and rifamycin-associated determinants in 35/513 (6.8%, 95% CI 4.9-9.3) (Table 2; Figure 2).

Table 2: Prevalence of the seven primary genomic AMR classes in ORTHO_STRICT.

AMR class	n/N	%	95% CI (%)
Macrolide	205/513	40.0	35.8-44.3
Methicillin	195/513	38.0	33.9-42.3
Lincosamide	177/513	34.5	30.5-38.7
Aminoglycoside	167/513	32.6	28.6-36.7
Quinolone	166/513	32.4	28.5-36.5
Tetracycline	49/513	9.6	7.3-12.4
Rifamycin	35/513	6.8	4.9-9.3

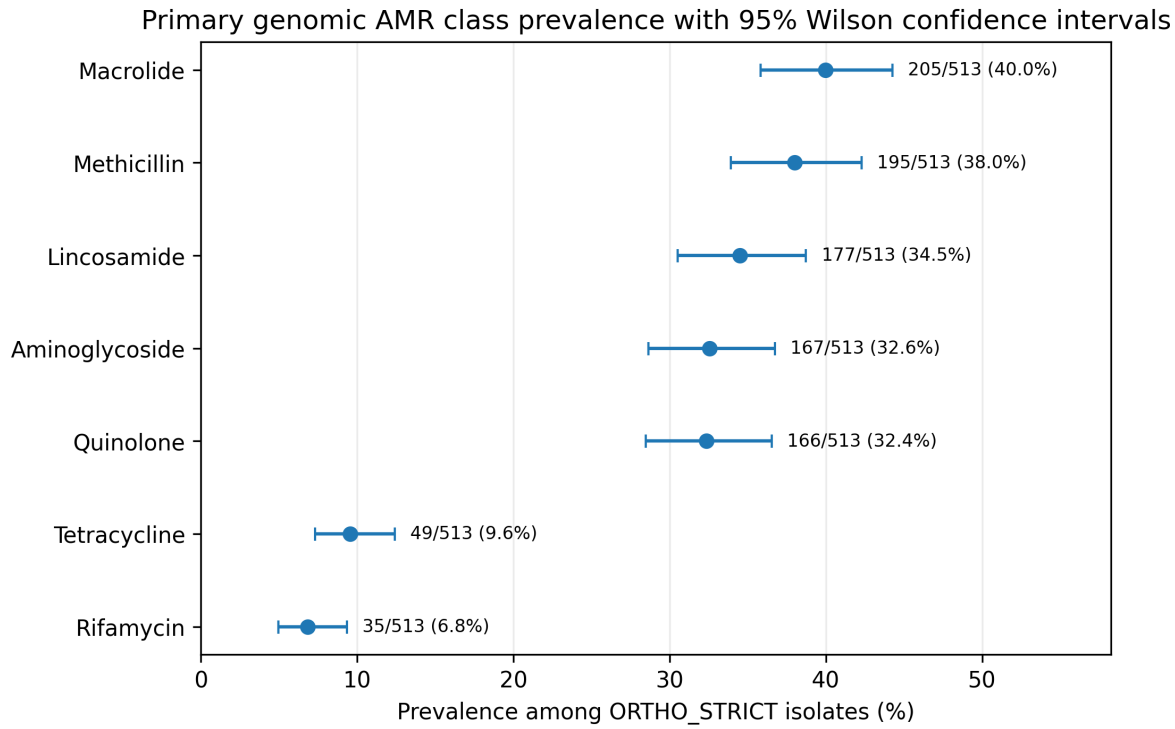


Figure 2: Prevalence of the seven prespecified primary genomic AMR classes among 513 ORTHO_STRICT isolates. Error bars are 95% Wilson confidence intervals.

Resistance architecture was heterogeneous. A total of 223/513 isolates (43.5%) carried none of the seven primary AMR classes. The most frequent multi-class profile comprised concurrent methicillin, macrolide, lincosamide, aminoglycoside and quinolone markers (51/513, 9.9%), while 21/513 isolates (4.1%) carried markers spanning all seven primary classes (Figure 3).

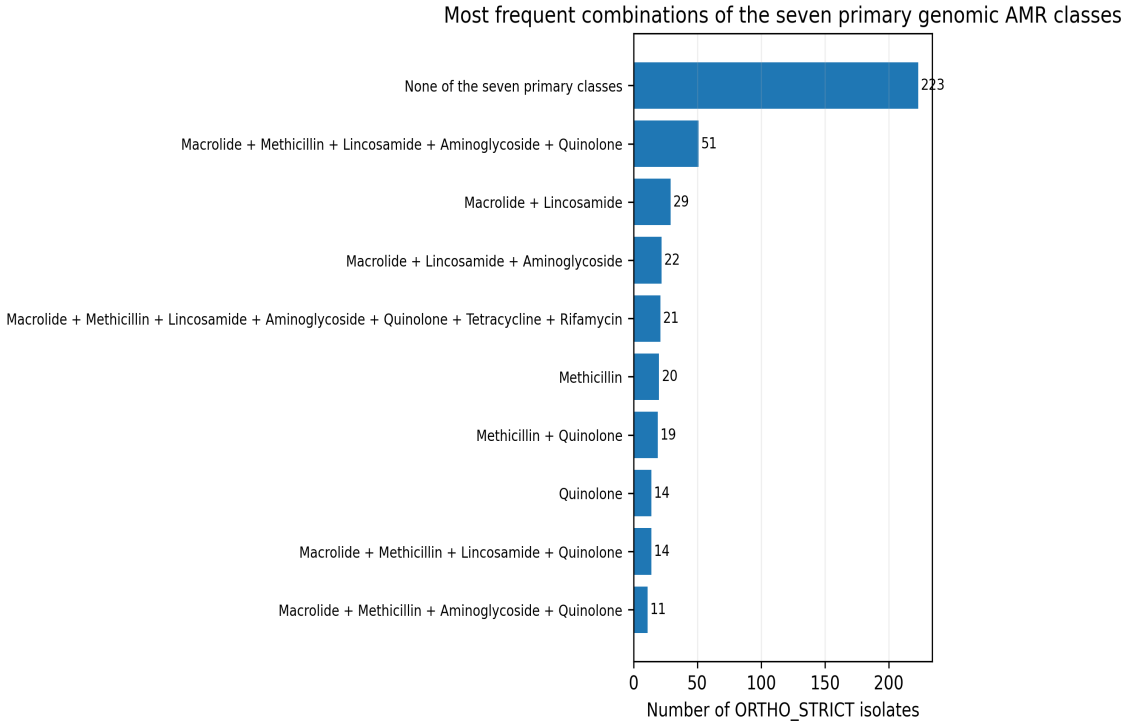


Figure 3: Ten most frequent combinations of the seven primary genomic AMR classes among ORTHO_STRICT isolates. "None" indicates absence of all seven prespecified class indicators, not absence of every possible AMR determinant.

Orthopaedic source-group AMR structure: Among the 495 isolates in the three adequately represented orthopaedic source groups (bone/osteomyelitis, joint/synovial and prosthesis/PJI-associated source), the first two PCoA axes accounted for 60.51% and 15.35% of positive-eigenvalue variation in the seven-class Jaccard distance matrix. The unstratified PERMANOVA yielded pseudo-F=12.040, permutation p=0.0005 and R2=0.0467. However, dispersion differed profoundly among groups (distance-to-centroid ANOVA p=3.28 x 10^-27), so the unstratified result could not be interpreted as robust evidence of centroid separation. Consistent with that caution, the BioProject-stratified PERMANOVA restricted to 96 isolates from seven multi-source BioProjects was null (p=0.671).

Descriptively, methicillin-associated determinants occurred in 121/289 bone isolates (41.9%), 62/147 joint isolates (42.2%) and 3/59 prosthesis-associated isolates (5.1%). Corresponding macrolide prevalences were 39.8%, 50.3% and 13.6%, and quinolone prevalences were 30.8%, 43.5% and 11.9%, respectively. These source-group distributions were interpreted alongside the project-stratified analyses.

Resistance-class burden: In the seven-primary-class burden analysis, 42.9% of bone/osteomyelitis isolates, 32.0% of joint/synovial isolates and 74.6% of prosthesis/PJI-associated-source isolates carried none of the seven primary classes.

In the year- and continent-adjusted BioProject-clustered Poisson GEE (N=479; 90 BioProjects), the joint-versus-bone incidence-rate ratio (IRR) was 0.835 (95% CI 0.672-1.038; p=0.105), whereas the prosthesis-versus-bone IRR was 0.368 (95% CI 0.155-0.876; p=0.0238); the overall subgroup Wald test was p=0.00323. Because this GEE primarily accounted for within-project dependence rather than all between-project confounding, the source-burden inference was treated as exploratory. The multi-source project sensitivity contained 96 isolates from seven projects, with only one prosthesis-associated isolate contributing to within-project overlap, so the prosthesis contrast could not be robustly confirmed.

Temporal and geographic patterns: Restricted cubic spline models for the better represented 2000-2025 period included 493 isolates from 95 BioProjects. No nonlinear association with collection year was detected for methicillin (p=0.942), macrolide (p=0.852), lincosamide (p=0.649), aminoglycoside (p=0.923) or quinolone (p=0.976). Among isolates with continent assignment, 265 were from Europe, 121 from North America, 55 from South America, 33 from Asia and 25 from Oceania. Exploratory year-adjusted models with BioProject-clustered covariance identified multiple continent-level contrasts that remained significant after Benjamini-Hochberg correction for several primary AMR classes; sparse contrasts were suppressed according to the prespecified event-cell rule. Because geography and contributing BioProject were strongly entangled and clustered covariance does not remove between-project confounding, these contrasts were not interpreted as regional prevalence differences. Full effect estimates and sparse-cell flags are reported in Supplementary Table S1.

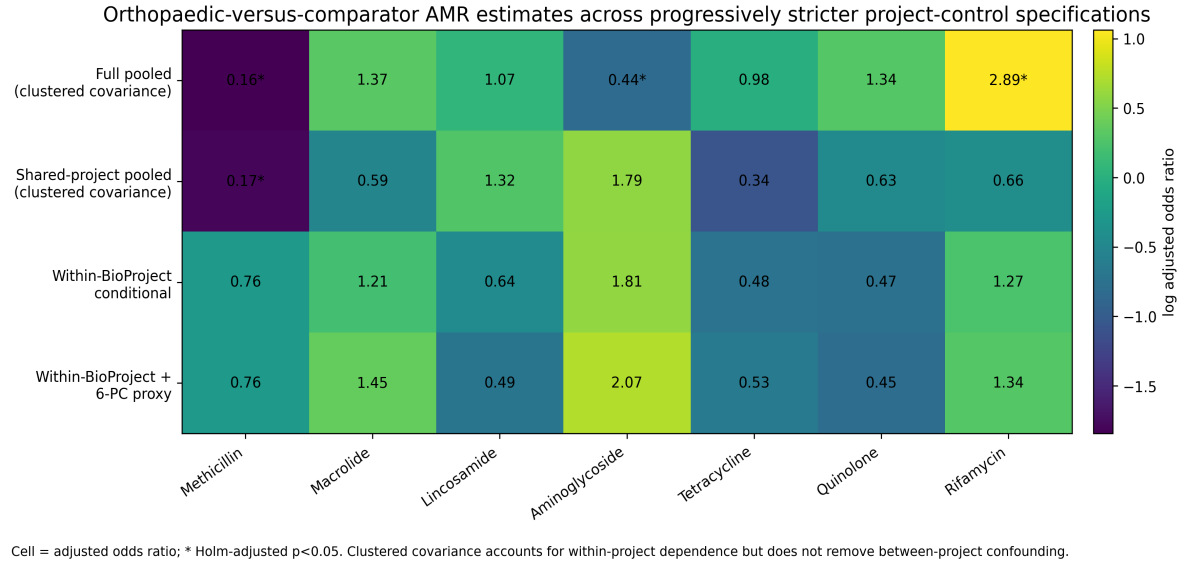
Orthopaedic-versus-comparator analysis: In the pooled complete-case joint multivariable model (N=9,561; 493 orthopaedic and 9,068 comparator isolates across 799 BioProjects), simultaneous adjustment for the seven primary AMR classes, collection year and continent with BioProject-clustered covariance identified three Holm-significant AMR coefficients: methicillin (aOR 0.158, 95% CI 0.069-0.362; Holm p=8.77 x 10^-5), aminoglycoside (aOR 0.436, 95% CI 0.282-0.674; Holm p=0.00112) and rifamycin (aOR 2.892, 95% CI 1.389-6.019; Holm p=0.0226). Macrolide, lincosamide, tetracycline and quinolone coefficients were not Holm-significant. Restriction to shared BioProjects yielded 1,399 complete cases (127 orthopaedic and 1,272 comparator isolates across 25 BioProjects); in this pooled shared-project model methicillin remained Holm-significant (aOR 0.167, 95% CI 0.058-0.479; Holm p=0.00610), whereas the other six classes did not. These pooled models accounted for correlated observations but remained susceptible to between-project confounding.

The stricter conditional logistic model used only within-BioProject orthopaedic-versus-comparator contrasts and included 1,414 isolates (127 orthopaedic and 1,287 comparator isolates) across the same 25 shared BioProjects. Different complete-case requirements explain the slightly larger denominator than in the pooled shared-project model. In the joint seven-class-plus-year model, no primary AMR class remained significant after Holm correction. Methicillin had an aOR of 0.760 (95% CI 0.289-2.000; p=0.579; Holm p=1.000); quinolone had an aOR of 0.473 (95% CI 0.204-1.096; p=0.0808; Holm p=0.566); and aminoglycoside had an aOR of 1.813 (95% CI 0.862-3.814; p=0.117; Holm p=0.700) (Table 3; Figure 4).

Table 3: Project-aware and lineage-proxy joint AMR-class estimates.

AMR class	Full multivariable aOR (95% CI); Holm p	Within-BioProject aOR (95% CI); Holm p	Within-BioProject + 6-PC aOR (95% CI); Holm p
Methicillin	0.158 (0.069-0.362); <0.001	0.760 (0.289-2.000); 1.000	0.759 (0.272-2.116); 1.000
Macrolide	1.370 (0.722-2.597); 1.000	1.214 (0.455-3.239); 1.000	1.452 (0.519-4.069); 1.000
Lincosamide	1.072 (0.573-2.007); 1.000	0.640 (0.256-1.599); 1.000	0.492 (0.178-1.359); 0.856
Aminoglycoside	0.436 (0.282-0.674); 0.001	1.813 (0.862-3.814); 0.700	2.072 (0.956-4.490); 0.454
Tetracycline	0.985 (0.607-1.598); 1.000	0.476 (0.178-1.271); 0.700	0.525 (0.192-1.438); 0.856
Quinolone	1.342 (0.765-2.354); 1.000	0.473 (0.204-1.096); 0.566	0.446 (0.188-1.059); 0.454
Rifamycin	2.892 (1.389-6.019); 0.023	1.270 (0.297-5.432); 1.000	1.339 (0.305-5.872); 1.000

Note: The full model used N=9,561; the strict within-BioProject and six-PC joint models used N=1,414 across 25 shared BioProjects. Six-PC estimates represent lineage-correlated accessory-genome background adjustment and are not direct MLST/clonal-complex adjustment.



Cell = adjusted odds ratio; * Holm-adjusted p<0.05. Clustered covariance accounts for within-project dependence but does not remove between-project confounding.

Figure 4: Change in orthopaedic-versus-comparator AMR estimates across progressively stricter project-control specifications.

Cells show adjusted odds ratios; asterisks denote Holm-adjusted p<0.05. BioProject-clustered covariance addresses within-project dependence but does not by itself remove between-project confounding; the conditional rows identify effects from within-BioProject contrasts. The final row additionally adjusts for six lineage-correlated accessory-genome principal components.

Within-BioProject and lineage-proxy robustness analyses: Direct sequence-type labels were recoverable from submitter-entered metadata for 173 isolates (25 orthopaedic and 148 comparator isolates), which was insufficient for full-cohort MLST *adjustment. The post hoc population-structure sensitivity therefore used 55 non-AMR virulence/stress presence-absence features. Six principal components explained 63.83% of accessory-genome variation. In the explicit-ST check subset (151 isolates across ST1, ST398, ST5, ST72 and ST630), the 10-PC space strongly separated sequence types (pseudo-F=189.701; R2=0.8386; 4,999-permutation p=0.0002) and achieved a five-fold cross-validated balanced accuracy of 0.96 (SD 0.055), supporting—but not proving—their use as lineage-correlated covariates.*

Within the 1,414-isolate, 25-BioProject strict shared-project cohort, adding the six accessory-genome PCs to collection year did not improve source discrimination (likelihood-ratio chi-square=1.726, df=6, p=0.943).

In single-class conditional models adjusted for year plus six PCs, no AMR class was Holm-significant; quinolone was nominally associated (aOR 0.446, 95% CI 0.219-0.909; p=0.0264) but did not survive Holm correction (p=0.184). The six-PC joint seven-class model likewise contained no Holm-significant AMR coefficient. The 10-PC sensitivity produced the same corrected conclusion; in the joint model aminoglycoside was nominally associated (aOR 2.233, 95% CI 1.014-4.915; p=0.0461) but not after Holm correction (p=0.323) (Table 3; Figure 5).

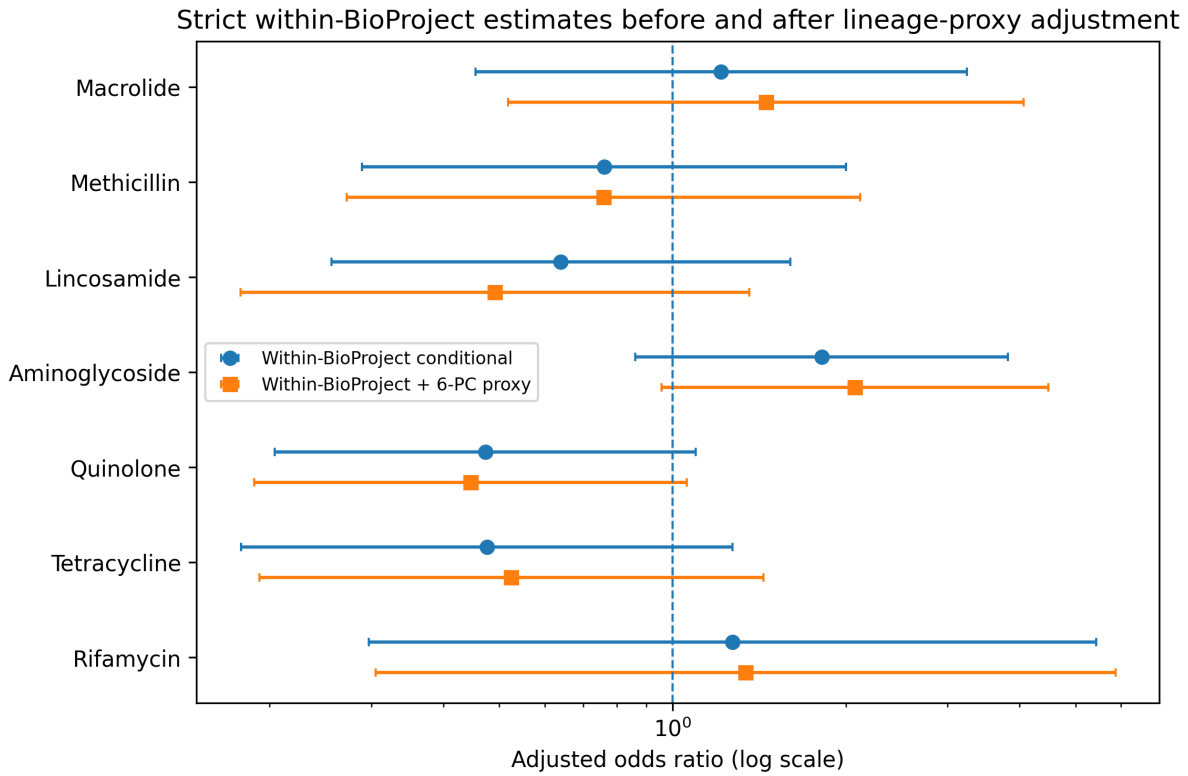


Figure 5: Strict within-BioProject joint-model estimates before and after six-PC lineage-proxy adjustment. Horizontal bars are 95% confidence intervals; the vertical line denotes an adjusted odds ratio of 1.0. No primary AMR class remained Holm-significant in either model.

The within-project bone-versus-joint analysis was limited to 71 isolates from six BioProjects (37 joint and 34 bone) and yielded no Holm-significant primary-class association. The rifamycin estimate was numerically unstable in the lineage-proxy sensitivity, and prosthesis-associated isolates had insufficient within-project overlap for a stable lineage-proxy source comparison. Primary-class prevalence was otherwise stable to AMRFinderPlus mode and source-definition sensitivity: restricting to the 408 COMBINED isolates changed prevalence by no more than 3.63 percentage points, and ORTHO_EXPANDED (N=535) changed primary-class prevalence by less than 1 percentage point.

Secondary AMR and AST description: Secondary genomic AMR annotation classes were not part of the seven prespecified primary-class hypothesis family and were not subjected to source-comparison inference in the main manuscript. Their prevalence, including fosfomycin-, beta-lactamase-, trimethoprim- and other lower-frequency annotations, is provided in Supplementary Table S2 to avoid conflating exploratory genomic annotations with phenotypic susceptibility. Only four ORTHO_STRICT BioSamples linked to submitted phenotypic AST data, contributing 44 isolate-antimicrobial test rows. The paired denominator was therefore retained for descriptive reporting only; genotype-phenotype sensitivity, specificity, positive predictive value and negative predictive value were not estimated.

Discussion

Three findings define the present analysis. First, publicly deposited orthopaedic-associated *Staphylococcus aureus* showed a substantial but heterogeneous genomic antimicrobial-resistance (AMR) burden: markers associated with macrolide and methicillin resistance were present in approximately two-fifths of isolates, while 43.5% carried none of the seven prespecified primary AMR classes and 9.9% carried the five-class methicillin–macrolide–lincosamide–aminoglycoside–quinolone combination. Second, apparent differences between orthopaedic specimen-source groups and between orthopaedic and non-orthopaedic clinical isolates were highly sensitive to the structure of the contributing sequencing projects. The unstratified source-group PERMANOVA was significant, but explained only 4.7% of the distance variation, occurred in the presence of marked dispersion heterogeneity, and was not reproduced in the BioProject-stratified sensitivity analysis; it therefore cannot be read as robust evidence of source-specific centroid separation. Third, once orthopaedic and comparator isolates were contrasted within the same

BioProjects, none of the seven primary AMR classes remained significant after multiplicity correction; a post hoc lineage-correlated accessory-genome sensitivity did not restore a robust orthopaedic-specific AMR signature. The central contribution is therefore not a claim of unusually high resistance in orthopaedic infection, but evidence that repository composition can materially influence clinical-source AMR associations.

The prevalence estimates provide clinically relevant context, but they are genomic determinant frequencies rather than an antibiogram. A 2026 systematic review and meta-analysis of musculoskeletal infections reported substantial between-study heterogeneity in *S. aureus* resistance, reinforcing variation by geography, infection type, period and sampling frame. [1] Recent orthopaedic whole-genome sequencing from rural China similarly described diverse resistance profiles across 80 isolates and multiple sequence types.4 Our finding of methicillin-associated determinants in 38.0% of isolates is therefore contextual surveillance information, not a population prevalence estimate. Genomic AMR annotation and phenotypic susceptibility testing answer related but non-equivalent questions: a detected gene or resistance-associated mutation indicates a genomic mechanism or prediction, whereas clinical susceptibility also depends on expression, genotype–phenotype concordance, testing method and breakpoints. Because paired orthopaedic AST data were sparse, these genomic frequencies cannot support empirical-treatment recommendations.

The resistance architecture was also heterogeneous. Nearly half of ORTHO_STRICT isolates lacked all seven primary class markers, while a smaller subset accumulated multiple major classes. This is consistent with the diversity reported in orthopaedic *S. aureus* cohorts. In a prospective study of orthopaedic device-related infections, 94% of invasive isolates were methicillin-susceptible, yet poor outcomes still occurred among MSSA lineages, illustrating that pathogenic success cannot be reduced to methicillin resistance alone. [17] Whole-genome analysis of osteomyelitis isolates from Italy likewise identified several common clonal complexes and variable virulence/resistance repertoires without a single genetic trait explaining osteomyelitis-associated success. [18] Prosthetic-joint infection isolates have also failed to form a unique PJI-specific phylogenetic cluster when compared with commensal controls. [5] These observations support describing the present data as a spectrum of multi-class genomic AMR profiles rather than a single "orthopaedic resistance phenotype," and argue against equating the seven-class count with phenotypic multidrug-resistance definitions. The strongest methodological result emerged as project structure was progressively controlled. At repository level, bone, joint and prosthesis-associated source groups showed an unstratified difference in the seven-class distance matrix, but the small R2, profound dispersion heterogeneity and null BioProject-stratified analysis prevented interpretation as a robust source-specific multivariate profile. A similar pattern occurred in orthopaedic-versus-comparator models. In the pooled multivariable model with BioProject-clustered covariance, methicillin, aminoglycoside and rifamycin coefficients were Holm-significant, and methicillin remained significant after restricting the pooled analysis to shared projects. Clustered covariance, however, addresses correlation rather than unmeasured between-project confounding. In the stricter conditional logistic analysis, using only within-BioProject contrasts across 1,414 isolates from 25 shared projects, none of the seven primary classes remained Holm-significant. For methicillin, the adjusted odds ratio moved from 0.158 in the pooled model to 0.760 within projects, with a confidence interval crossing unity; other classes also changed in magnitude and, in some cases, direction. This does not prove that the pooled associations were artefactual; it shows that the inference was not stable to stronger control of the data-generating structure.

That instability is plausible in a public genomic repository. Previous NCBI Pathogen Detection analyses documented source- and geography-associated variation in *S. aureus* resistance genes across large One Health datasets, demonstrating both the platform's scale and uneven submission composition. [9,10] More generally, bacterial genomic datasets are not collections of exchangeable independent observations. Yu and colleagues showed across more than 24,000 genomes from five pathogens that biased sampling coupled to population structure can seriously confound AMR prediction, and that larger sample size does not automatically solve the problem. [11] Earle and colleagues likewise demonstrated the need for explicit control of lineage in bacterial association studies. [12] BioProject can pragmatically capture several linked sources of heterogeneity—sampling purpose, institution, geography, period, local lineage composition and sequencing/submission workflow. The attenuation seen here therefore supports testing clinical-source associations within, or explicitly accounting for, contributing project structure rather than relying only on pooled repository comparisons.

The post hoc lineage-proxy analysis strengthens this interpretation while defining an important boundary. Direct sequence-type information was too sparse for full-cohort MLST adjustment, so population structure was represented using principal components from 55 non-AMR

virulence/stress features. These components were strongly lineage-correlated in the limited explicit-ST subset: the 10-PC space showed marked between-ST separation (R^2=0.839) and cross-validated balanced accuracy of 0.96 across five adequately represented ST groups. Nevertheless, adding the primary six-PC block to the within-BioProject source model did not improve orthopaedic-versus-comparator discrimination (likelihood-ratio p=0.943), and no primary AMR class became Holm-significant in six-PC or 10-PC models. Orthopaedic genomic studies provide a useful parallel: Noone et al. showed the importance of interpreting implant-associated signals against lineage structure, while Wildeman et al. found no PJI-specific genomic cluster. [5,6] Because the proxy uses accessory features rather than direct phylogeny, it should be understood as a sensitivity analysis only; it does not establish clonal-complex independence or replace direct MLST/cgMLST/core-genome analysis.

Temporal and geographic findings require similarly restrained interpretation. Restricted cubic-spline analyses for 2000–2025 detected no robust nonlinear temporal association for methicillin, macrolide, lincosamide, aminoglycoside or quinolone markers. This is compatible with some longitudinal orthopaedic literature: Duployez et al. found staphylococcal susceptibility broadly stable over 2010–2019, with decreases in rifampicin and fluoroquinolone resistance, [20] and Chandler et al. reported no significant decade-long trend in the proportion of surgically treated PJI caused by resistant organisms. [21] Other local series have documented increasing methicillin resistance or shifts in selected drug classes. [22,23] These trajectories arise from different hospitals, case mixes, organisms, phenotypic methods and time windows. Our repository-based null therefore does not imply that resistance is globally static. Likewise, significant continent-level contrasts among deposited isolates describe database composition, not national or regional population prevalence.

One exploratory observation deserves particular caution. Prosthesis/PJI-associated source isolates had a lower seven-class burden than bone isolates in the BioProject-clustered GEE, but that model primarily addressed within-project dependence and the multi-source project sensitivity contained only one prosthesis-associated isolate, precluding robust within-project confirmation. A lower AMR burden is biologically conceivable because successful device infection may depend on traits such as adhesion, biofilm formation, and persistence and host–implant interactions rather than on broad resistance alone. The prospective orthopaedic device study by Post et al. is relevant in showing that predominantly methicillin-susceptible lineages can still be associated with poor outcomes, [17] while PJI genomic studies have emphasized heterogeneous lineages and multifactorial pathogenesis. [5,19] However, the present metadata do not measure biofilm phenotype, treatment exposure or clinical outcome. The prosthesis-associated signal should therefore remain hypothesis-generating and should not be used to infer a distinct biological or therapeutic phenotype.

Several design features strengthen the study. The primary cohort used a high-specificity, rule-based orthopaedic specimen-source definition and was complemented by an expanded-source sensitivity cohort. A large, deliberately restricted clinical comparator avoided blood, generic wound/tissue, surveillance and ambiguous-source isolates that could contaminate the contrast. This restriction improves specificity but also narrows the comparator case mix and therefore limits generalisability to all non-orthopaedic *S. aureus* infections. The analytical sequence moved from pooled comparisons to shared-project analyses, strict within-BioProject conditional models, leave-one-project-out checks and an empirically supported lineage-correlated population-structure sensitivity. Multiplicity correction was applied to the seven primary classes, and nominal associations that failed correction were not promoted as findings. Primary prevalence estimates were also stable to the expanded orthopaedic definition and to restriction to AMRFinderPlus COMBINED analyses. Together, these features make the negative robustness results an analytical outcome rather than a failure to detect significance.

The limitations determine the scope of inference. NCBI Pathogen Detection is a convenience repository assembled from heterogeneous research and surveillance submissions, so neither cohort represents population prevalence. The analytical unit is the deposited isolate, not the patient; patient-level linkage was unavailable and residual non-independence from repeat sampling cannot be fully excluded. Orthopaedic classification relied on free-text metadata and, although strict rules prioritised specificity and strict-versus-expanded sensitivity was reassuring, independent clinician adjudication has not yet been completed and remains an important external-validity limitation. Paired phenotypic AST was too sparse for meaningful genotype–phenotype accuracy analysis. Direct MLST, clonal-complex or phylogenetic calls were unavailable across the full cohort, so the post hoc accessory-genome PCs remain a lineage-correlated proxy and may also capture non-lineage biological variation. Geographic models were limited by project–geography entanglement, and within-project overlap was limited for some orthopaedic subgroups, especially prosthesis-associated isolates. Finally, the deliberately narrow

comparator reduces source misclassification but may not represent the full spectrum of non-orthopaedic clinical S. aureus.

Taken together, the findings support project-aware secondary bacterial genomic epidemiology. Public repositories such as NCBI Pathogen Detection can describe genomic AMR breadth and generate hypotheses at a scale difficult to achieve prospectively, but pooled clinical-niche comparisons can overstate apparent differences when project, geography and lineage are unevenly distributed. For orthopaedic S. aureus, the robust message is dual: substantial and diverse genomic resistance determinants are present, yet no major AMR class showed a stable orthopaedic-specific association after strict within-BioProject and lineage-proxy adjustment. Future studies should combine direct sequence typing or phylogeny with prospectively curated specimen definitions, patient-level clinical data and paired phenotypic testing. Public genomic surveillance should therefore complement—not replace—local antibiograms and clinically linked prospective surveillance.

Conclusion

Among 513 high-confidence orthopaedic-associated *Staphylococcus aureus* isolates in NCBI Pathogen Detection, genomic determinants associated with several major antimicrobial-resistance classes were common, but resistance architecture was heterogeneous and no single dominant orthopaedic profile emerged. Apparent differences across orthopaedic specimen-source groups and between orthopaedic and other clinical isolates were substantially weakened, became imprecise or changed direction when comparisons were restricted within contributing BioProjects; a post hoc lineage-correlated accessory-genome sensitivity did not materially change that corrected conclusion.

These findings support the use of large public genomic repositories for descriptive AMR surveillance and hypothesis generation, while showing that sequencing-project composition and bacterial population structure must be considered before interpreting pooled clinical-source associations as biological differences. Because the repository is a non-probability sample and paired phenotypic susceptibility data were sparse, these genomic estimates should complement rather than replace local antibiograms, clinically linked surveillance and prospective genomic studies with direct lineage assignment and patient-level metadata.

Ethics Statement

This secondary analysis used only publicly accessible, de-identified microbial genomic and contextual metadata from NCBI Pathogen Detection. No participants were recruited, no new specimens were collected, no clinical intervention occurred, no private patient records were accessed, and no attempt was made to re-identify individuals.

Data Availability

Source data were obtained from the frozen NCBI Pathogen Detection *Staphylococcus aureus* release PDG000000073.1274, released 3 September 2026 and extracted 4 September 2026. Frozen analysis datasets, cohort and exclusion logs, statistical scripts, model outputs, figures, random seeds and provenance records are retained and can be supplied to the journal editor for verification if requested.

References

1. Wang B, Li M, Yu J, Jiang F, Wang C, Hu Y, et al. Global prevalence and antimicrobial resistance of *Staphylococcus aureus* isolates from patients with musculoskeletal infections: a systematic review and meta-analysis. Int J Surg 2026;112(2):4981-4989. doi:10.1097/JS9.0000000000003862.
2. Morgenstern M, Erichsen C, Militz M, Xie Z, Peng J, Stannard J, et al. The AO trauma CPP bone infection registry: Epidemiology and outcomes of *Staphylococcus aureus* bone infection. J Orthop Res 2021;39(1):136-146. doi:10.1002/jor.24804.
3. Bouiller K, David MZ. *Staphylococcus aureus* Genomic Analysis and Outcomes in Patients with Bone and Joint Infections: A Systematic Review. Int J Mol Sci 2023;24(4):3234. doi:10.3390/ijms24043234.
4. Wang Y, Wang Y, Liu C, Zhao D, Cui Y, Guan X, et al. Genomic and clinical characteristics of *Staphylococcus aureus* isolates associated with orthopedic infections from a rural hospital, Qingdao, China. Microbiol Spectr 2026;14(5):e03692-25. doi:10.1128/spectrum.03692-25.
5. Wildeman P, Tevell S, Eriksson C, Campillay Lagos A, Söderquist B, Stenmark B. Genomic characterization and outcome of prosthetic joint infections caused by Staphylococcus aureus. Sci Rep 2020;10(1):5938. doi:10.1038/s41598-020-62751-z.
6. Noone JC, Stegger M, Lilje B, Stavem K, Helmersen K, Skråmm I, et al. Molecular characteristics of *Staphylococcus aureus* associated prosthetic joint infections after hip fractures treated with hemiarthroplasty: a retrospective genome-wide association study. Sci Rep 2020;10(1):16553. doi:10.1038/s41598-020-73736-3.
7. National Center for Biotechnology Information. Pathogen Detection Help Document [Internet]. Bethesda (MD): National Library of Medicine; [cited 2026 Sep 6].

Available from: https://www.ncbi.nlm.nih.gov/pathogens/pathogens_help/.
8. Feldgarden M, Brover V, Fedorov B, Haft DH, Prasad AB, Klimke W. Curation of the AMRFinderPlus databases: applications, functionality and impact. Microb Genom 2022;8(6): mgen000832. doi:10.1099/mgen.0.00083 2.
9. Zaghen F, Sora VM, Meroni G, Laterza G, Martino PA, Soggiu A, et al. Epidemiology of Antimicrobial Resistance Genes in *Staphyloccocus aureus* Isolates from a Public Database in a One Health Perspective—Sample Characteristics and Isolates' Sources. Antibiotics (Basel) 2023;12(7):1225. doi:10.3390/antibiotics12071225.
10. Zaghen F, Sora VM, Meroni G, Laterza G, Martino PA, Soggiu A, et al. Epidemiology of Antimicrobial Resistance Genes in *Staphylococcus aureus* Isolates from a Public Database from a One Health Perspective—Sample Origin and Geographical Distribution of Isolates. Antibiotics (Basel) 2023;12(12): 1654. doi:10.3390/antibiotics12121654.
11. Yu Y, Wheeler NE, Barquist L. Biased sampling driven by bacterial population structure confounds machine learning prediction of antimicrobial resistance. PLoS Biol 2025;23(12):e3003539. doi:10.1371/journal.pbio.3003539.
12. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol 2016;1:16041. doi:10.1038/nmicrobiol.2016.41.
13. Wohl S, Lee EC, DiPrete BL, Lessler J. Sample size calculations for pathogen variant surveillance in the presence of biological and systematic biases. Cell Rep Med 2023;4(5): 101022. doi:10.1016/j.xcrm.2023.101022.
14. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 2021;11:12728. doi:10.1038/s41598-021-91456-0.
15. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (Strobe) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008; 61(4):344-349. doi:10.1016/j.jclinepi.2007.11.008.
16. Field N, Cohen T, Struelens MJ, Palm D, Cookson B, Glynn JR, et al. Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. Lancet Infect Dis 2014;14(4):341-352. doi:10.1016/S1473-3099(13)70324-4.
17. Post V, Pascoe B, Hitchings MD, Erichsen C, Fischer J, Morgenstern M, et al. Methicillin-sensitive *Staphylococcus aureus* lineages contribute towards poor patient outcomes in orthopaedic device-related infections. Microb Genom 2025;11(4):001390. doi:10.1099/mgen.0.001390.
18. Pimentel de Araujo F, Pirolo M, Monaco M, Del Grosso M, Ambretti S, Lombardo D, et al. Virulence determinants in *Staphylococcus aureus* clones causing osteomyelitis in Italy. Front Microbiol 2022;13:846167. doi:10.3389/fmicb.2022.846167.
19. Muñoz-Gallego I, Viedma E, Esteban J, Mancheño-Losa M, García-Cañete J, Blanco-García A, et al. Genotypic and phenotypic characteristics of *Staphylococcus aureus* prosthetic joint infections: insight on the pathogenesis and prognosis of a multicenter prospective cohort. Open Forum Infect Dis 2020;7(9):ofaa344. doi:10.1093/ofid/ofaa344.
20. Duployez C, Millière L, Senneville E, Piantoni L, Migaud H, Wallet F, Loïez C. Evolution of antibiotic susceptibility profiles of staphylococci from osteoarticular infections: a 10-year retrospective study. Orthop Traumatol Surg Res 2023;109(8):103512. doi:10.1016/j.otsr.2022.103512.
21. Chandler CC, Graham SD, Hietpas KT, Fehring TK, Otero JE. Trends in antibiotic-resistant bacteria in surgically treated periprosthetic joint infection: a decade of experience at a regional referral center. J Arthroplasty 2026;41(7):2146-2153. doi:10.1016/j.arth.2025.10.119.
22. Gürler M, Kırca F, Dinç B. Septic arthritis and prosthetic joint infections: microbial spectrum and evolving resistance patterns. J Bone Jt Infect 2026;11(1):123-137. doi:10.5194/jbji-11-123-2026.
23. Ait Said L, Debbagh F, Errahmani MY, Elghazouani Z, Abouankira M, Gharmoud M, et al. Temporal changes in microbial etiology and antibiotic resistance profiles of osteoarticular infections: an 11-year retrospective analysis. Cureus 2025;17(10): e95572. doi:10.7759/cureus.95572.