Domain-Aware Versus Machine Learning Imputation for Sparse Antimicrobial Susceptibility Data

Authors: *Fredrick Mutisya,[1] Taïoh Yokoyama,[1] Sana Boujaafar,[1] Cyprien de Turckheim,[1] Mathieu Raad[1]

1. SmartBiotic, Strasbourg, France
*Correspondence to fredrick.mutisya@smartbiotic.ai

Disclosure: Raad, de Turckheim, and Yokoyama have received support from SmartBiotic for the present manuscript. Mutisya, de Turckheim, Raad, and Yokoyama have received consulting fees from SmartBiotic. SmartBiotic is a software platform analysing local bacterial ecology of hospitals to develop tailored antibiotic therapy recommendations. Boujaafar has declared no conflicts of interest.

Acknowledgements: The authors would like to thank Pfizer and Vivli for open access to the Atlas dataset.

Keywords: Antimicrobial susceptibility, imputation, machine learning (ML), sparse data.

Citation: EMJ Microbiol Infect Dis. 2026;7[1]:36-37. https://doi.org/10.33590/emjmicrobiolinfect-dis/XXJQ9810

BACKGROUND AND AIMS

Antimicrobial susceptibility testing datasets frequently miss data in a structured way due to selecting testing practices. Handling these gaps with generic statistical imputation may violate well-established microbiological rules such as intrinsic resistance or non-reportable combinations. This study aims to compare domain-aware completion based on established microbiology rules with several machine learning (ML)-based imputation strategies under controlled conditions.[1]

MATERIALS AND METHODS

The authors evaluated imputation strategies using the Atlas dataset 2022 (Pfizer, New York, USA) for levofloxacin, meropenem, and gentamicin, selected to represent different drug classes and resistance patterns. SmartBiotic's (Montreal, Canada) rule engine derived from the European Committee on Antimicrobial Susceptibility Testing (EUCAST) and Clinical and Laboratory Standards Institute (CLSI) expected resistance and susceptibility phenotypes was used for domain-aware completion. For validation, 20% of truly observed results were randomly masked per antibiotic, simulating missing completely at random conditions. Five reconstruction strategies of masked values were compared: global frequency imputation, ML with listwise deletion, ML with random undersampling, ML with synthetic minority oversampling technique (SMOTE), and a rule-based strategy. Outcomes were binarised (susceptible versus resistant/intermediate) and evaluated using accuracy, sensitivity, specificity, predictive values, F1 score, and Cohen's kappa.

RESULTS

The authors' inferred resistance rules added 75,730 new cells, mainly from intra-species inference (75,729 cells), while intrinsic resistance rules augmented all 55,549 rows with 502,847 additional cells (11.2%). In the masking experiment, the domain rule completion demonstrated high performance for levofloxacin (accuracy: 95%; sensitivity: 92%; specificity: 96%), meropenem (91%/97%/90%), and gentamicin (86%/57%/96%). With imbalance handling, ML with SMOTE oversampling achieved moderate sensitivity improvement (levofloxacin: 58%; meropenem: 92%; gentamicin: 54%) over unbalanced ML (43%/90%/41%). Random undersampling produced balanced profiles but lower overall performance (66–72% across metrics). Frequency imputation yielded 0% sensitivity across all antibiotics despite acceptable accuracy (69–79%).

CONCLUSION

These findings suggest that antimicrobial susceptibility testing missingness should

first be addressed as a microbiological problem and only then as a statistical one. Rule-based systems predict only when applicable, yielding high specificity but variable coverage-depending sensitivity. ML required imbalance correction for meaningful resistance detection, with SMOTE oversampling offering optimal compromise. These results, consistent with recent ML approaches in AMR surveillance, support hierarchical imputation with deterministic approach first, then imbalance-aware ML for residual gaps. Future validation work on these approaches should assess performance under real-world settings.

References

1. Mutisya F et al. Domain-aware versus machine learning imputation for sparse antimicrobial susceptibility data. Abstract O0401. ESCMID Global, 17-21 April, 2026.