UNIVERSITY OF CATANIA

Department of Electrical, Electronic and Computer Engineering
Ph.D. in "Systems, Energy, Computer and Telecommunications Engineering"

ENGINEERING METHODS FOR MODELING ANTIMICROBIAL RESISTANCE

Candidate:
Chiara Condorelli

Supervisors:
Prof. Mattia Frasca
Prof. Lucia Valentina Gambuzza
Prof. Vincenza Carchiolo

Cycle XXXVIII

Acknowledgments

The work presented in this thesis was made possible thanks to the support and collaboration of many people and institutions.

First of all, I would like to express my sincere gratitude to my supervisors, Prof. Mattia Frasca, Prof. Lucia Valentina Gambuzza, and Prof. Vincenza Carchiolo, for giving me the opportunity to begin this Ph.D. program and for their guidance, commitment, and continuous support throughout these three years.

I am also very grateful to Professor Jesús Gómez-Gardeñes of the University of Zaragoza, who welcomed me into his research group and offered me the opportunity to undertake a scientifically enriching and personally meaningful research stay abroad. My thanks go to all my colleagues at the University of Zaragoza for their kindness and support during my months there.

I also thank the colleagues from the Biometec Department of the University of Catania, with whom I collaborated during the development of part of this research, in particular Dr. E. Nicitra, Dr. N. Musso, Dr. D. Bongiorno, and Prof. S. Stefani, for their scientific contribution.

A heartfelt acknowledgment goes to all my colleagues from Lab-1 at the University of Catania: thank you for your support and for the working environment we shared. I am especially grateful to Alessandra and Cinzia, whose companionship and friendship have been invaluable throughout this journey.

I would also like to thank my friends for their constant support.

Finally, my deepest gratitude goes to Gabriel and to my family, whose unwavering support has allowed me to overcome every difficulty along the way.

This work was supported by the Italian Ministry of University and Research (MUR) under the PNRR Extended Partnership Initiative on Emerging Infectious Diseases (Project No. PE00000007, INF-ACT).

Contents

Chapter 1

Introduction

1.1 The AMR Problem

Antimicrobial resistance (AMR) is defined as the ability of microorganisms to survive or grow in the presence of antimicrobial drugs that would normally inhibit or kill them, and in particular, it refers to the ability of bacteria to survive after exposure to a specified concentration of antimicrobial substances [37, 3, 67]. Surviving antibiotic effects is a normal bacterial reaction, leading to the creation of a clone capable of resisting the antibiotic. Although resistance can be intrinsic, resulting from natural bacterial physiology, it is of particular concern when acquired through genetic mutations or horizontal gene transfer [58]. Resistant strains are classified by genus, species, and antibiotic resistance phenotype [35]. This phenotype is established by comparing the list of antibiotics active on the reference strain with those the tested strain is resistant to, representing acquired resistance. In the bacterial domain, where this thesis is focused, AMR often leads to multidrug-resistant (MDR), extensively drug-resistant (XDR), or pan-drug-resistant (PDR) phenotypes [74] that severely restrict therapeutic options.

The global burden of AMR is substantial and has been recognized by the World Health Organization (WHO) as one of the major threats to human health in the 21st century [26, 84]. A landmark study [80] estimated that in 2019, 4.95 million deaths were associated with bacterial AMR, of which 1.27 million were directly attributable. More recent analyses [81] extended this

evidence across the period 1990–2021 and provided forecasts suggesting that the mortality burden of AMR will continue to rise in the coming decades. These figures, while based on the best available modeling approaches, remain subject to uncertainties due to incomplete surveillance data.

Clinicians have several reasons to be concerned about bacterial resistance, as resistant bacteria, particularly *Staphylococcus aureus*, *Enterococcus spp.*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*, are becoming increasingly prevalent in healthcare facilities [78, 39]: healthcare-associated infections (HAIs) serve as a reservoir and breeding ground for resistant bacteria [90, 65, 99]. These pathogens also appear in the 2024 revision of the WHO Priority Pathogens List [49], where carbapenem-resistant Gram-negative bacteria are classified as "critical" targets for research and development. The inability to cure infectious diseases with antibiotics raises concerns about the future of healthcare, leading to serious illnesses, prolonged hospital stays, rising healthcare costs, and treatment failures [24].

In addition to the cost in health and human lives, AMR is also an economic challenge. To have an idea, the Organisation for Economic Co-operation and Development (OECD) estimated that resistant infections generate annual costs of approximately USD 66 billion across 34 high-income countries [69], driven by longer hospital stays, higher medical expenses, and productivity losses.

These data highlight that AMR represents both a global health crisis and an economic challenge. For these reasons, new methodological approaches are needed to improve our ability to describe, predict, and mitigate AMR, particularly in hospital settings, where its consequences are most immediate and severe.

1.2 Motivations

While AMR is widely recognized as one of the major health threats of our time, most strategies to address it still rely on descriptive surveillance, antibiotic stewardship, and infection control measures, which are necessary but not sufficient. In fact, surveillance data are often fragmented, heterogeneous across hospitals and countries, and not easily generalizable. Traditional statistical approaches, although useful for trend monitoring, typically fail to

capture the nonlinear and multiscale dynamics that govern the emergence and spread of resistance within complex healthcare environments [55].

Recent advances in computational methods provide new opportunities to overcome these limitations. Machine Learning (ML) techniques have been increasingly applied to clinical and microbiological datasets to predict resistance patterns and guide empirical antibiotic therapy [97, 86]. In the hospital setting, predictive models have already demonstrated their potential to anticipate resistance trends based on antibiotic consumption and historical data [112]. At the same time, mathematical and population-based models offer a complementary perspective, simulating transmission dynamics of resistant pathogens across wards and healthcare networks [31]. However, systematic reviews highlight that these models often lack external validation, rely on oversimplified assumptions, and remain underutilized in clinical or policy decision-making [16].

Against this background, the motivation for this thesis is to contribute to bridging these gaps by developing and applying engineering methods to the study of AMR. In this context, the ability to predict AMR at the strain level is essential for supporting timely and informed clinical decision-making. Rapid and accurate prediction can improve empirical therapy, reduce the misuse of broad-spectrum antibiotics, and ultimately slow down the selective pressure driving resistance. Yet, predictive power alone is insufficient if the available data are incomplete or biased. Many genomic and phenotypic AMR datasets suffer from missing values due to heterogeneous laboratory protocols, selective testing, or incomplete sequencing. This motivates the inclusion of a dedicated analysis of data reconstruction, where ML is employed to impute missing values and enhance the robustness and interpretability of predictive models.

Beyond the data-driven perspective, a second motivation arises from the need to understand the mechanisms underlying the spread of resistance in healthcare environments. While ML can identify correlations, mechanistic population models capture causal dynamics-how transmission, mobility, and selective pressure interact to shape AMR patterns across wards. Developing such models allows us to explore intervention scenarios, estimate epidemic thresholds, and quantify the impact of control strategies.

In summary, the aim is to provide predictive and mechanistic insights

that go beyond descriptive statistics, thus supporting more effective antimicrobial stewardship, infection control interventions, and ultimately patient care.

1.3 Structure of the thesis

This thesis is organized into two main parts, each addressing a distinct but complementary aspect of the work conducted during the doctoral research. Together, they reflect the general goal of developing and applying engineering-based methodologies to the study of AMR in hospital environments.

Part I – Methods based on Machine Learning for AMR Prediction and Data Reconstruction

The first part focuses on data-driven approaches, exploring how ML techniques can be applied to predict AMR phenotypes and reconstruct incomplete biological datasets. It begins with an introduction to the fundamental principles of ML and their relevance in the biomedical context, followed by an overview of the main algorithms employed for classification and regression tasks. Subsequently, the datasets used throughout the research are described, including binary, continuous, and synthetic data derived from genomic and phenotypic sources. A dedicated section presents the evaluation metrics adopted to assess model performance, both for classification and for imputation tasks. The experimental results are then organized into two main applications: AMR prediction from genomic and phenotypic data, and reconstruction of missing values in biological datasets. Each application is examined through an analysis that includes specific preprocessing steps (such as correlation analysis, data balancing, and feature handling), the definition of model configurations, and a systematic comparison across multiple machine learning methods. For AMR prediction, the results highlight how different classifiers perform across heterogeneous datasets, while for missing data reconstruction, the analysis quantifies the robustness of various imputation strategies under two different types of missing mechanisms. This part concludes with a synthesis of the key findings and the methodological insights gained from the ML analyses.

Part II – Population Models

The second part of the thesis shifts from data-driven to mechanistic approaches, focusing on the mathematical modeling of AMR transmission in hospital settings. It introduces the conceptual background of population models, reviewing classical epidemic frameworks (such as SIS, SIR, and SEIR) and their extensions to metapopulation and structured systems. The theoretical framework of reproduction numbers and the Next Generation Matrix is then developed, providing the analytical tools used in subsequent chapters. Based on this formalism, a compartmental model, the XSR model, is proposed to describe AMR dynamics at both single-ward and multi-ward (metapopulation) levels. The analysis includes the derivation of equilibria, stability conditions, and critical thresholds of transmission parameters. Finally, the same theoretical principles are extended to a new more complex SIIS model, allowing the estimation of the basic and invasion reproduction numbers and the exploration of coexistence scenarios between sensitive and resistant strains. Numerical simulations, both deterministic and stochastic, are used to validate the theoretical results and to investigate the system's behavior under varying epidemiological conditions. The part concludes by highlighting the implications of these modeling approaches for understanding AMR persistence and control within hospital networks.

The thesis concludes with a general discussion that connects the results of the two parts, emphasizing how data-driven and theoretical modeling approaches can jointly enhance the understanding and management of AMR. By integrating ML and population dynamics, the work aims to provide methodological and conceptual contributions that may support future research and inform decision-making in clinical and public health contexts.

Part I

Methods based on Machine Learning for AMR Prediction and Data Reconstruction

Chapter 2

Introduction to Machine Learning Methods for AMR Prediction and Data Reconstruction

As mentioned in the previous chapter, AMR is widely recognized as an increasing threat to global public health. The inability to adequately treat infectious diseases with available antibiotics leads to severe consequences, including treatment failures, prolonged hospital admissions, increased healthcare costs, and ultimately higher morbidity and mortality [24]. The problem is particularly acute for vulnerable patients undergoing medical treatments that compromise the immune system, as well as for individuals with chronic conditions such as diabetes, asthma, rheumatoid arthritis, or cystic fibrosis [46, 70].

Timely and reliable identification of resistant bacterial strains is essential for guiding antibiotic prescription and reducing inappropriate drug use. However, traditional diagnostic methods, including culturing and antimicrobial susceptibility testing, remain time-consuming and resource-intensive. This limitation has motivated a growing interest in the use of ML techniques to provide rapid and accurate predictions of antimicrobial resistance. Recent studies highlight the effectiveness of ML algorithms in predicting AMR across multiple bacterial species and resistance mechanisms [66], as we will

discuss in more detail in the next section.

2.1 Machine Learning Approaches to AMR Prediction

ML algorithms have shown very good performance in predicting AMR mechanisms such as efflux pumps, target modifications, and enzymatic inactivation [79, 82]. Classical supervised learning techniques, including support vector machines, logistic regression, and random forests, have consistently demonstrated high predictive accuracy. More recently, deep learning models have expanded the methodological toolkit, offering the capacity to capture nonlinear dependencies and high-dimensional genomic features. When trained on whol genome sequencing (WGS) data, these models provide a time-efficient alternative to traditional laboratory diagnostics.

Beyond the classification of resistant versus susceptible strains, ML approaches have also been successfully used in broader areas of AMR research. For instance, they have been used to predict new resistance genes, antimicrobial peptides, and even candidate antibiotics, thus supporting both diagnostic applications and drug discovery pipelines [7, 105]. A comprehensive overview of ML-driven solutions to the AMR challenge can be found in [32].

Recent literature highlights attempts to use ML algorithms to improve prescribing practices, with promising results even when models are trained on relatively small datasets [83, 17, 36]. In particular, these models often outperform traditional logistic regression, underscoring the added value of ML in antibiotic management. Complementary research has also explored the application of ML to datasets obtained through mass spectrometry, further broadening the scope of predictive methodologies [116].

The potential of ML for AMR prediction has been demonstrated across several pathogens. Studies have targeted *Escherichia coli* to model resistance against multiple antibiotics [94, 79], while others have focused on *Staphylococcus aureus* [18, 115, 114] or *Streptococcus pneumoniae* [106]. In our work, we focus our attention in particular on *Klebsiella pneumoniae*. In the context of *Klebsiella pneumoniae*, ML analyses have been applied to predict phenotypic polymyxin resistance [72], to identify resistance features from single-nucleotide polymorphism (SNP) data [64], and to perform

multilabel classification with missing labels [107]. These studies provide evidence of the robustness and versatility of ML methods in handling various AMR-related tasks.

2.2 The Challenge of Missing Data in AMR Datasets

A persistent obstacle in AMR research is the problem of missing or incomplete data. AMR datasets often suffer from inconsistencies arising from incomplete sequencing, experimental errors, or heterogeneous reporting standards. Missing data can introduce substantial bias, degrade statistical analyses, and compromise the reliability of predictive models. Traditional strategies for handling missing data, such as listwise deletion or mean imputation, are simple but often inadequate, as they risk oversimplifying biological complexity and discarding valuable information [63].

More advanced statistical techniques, including Multiple Imputation by Chained Equations (MICE) [109, 8] and the Expectation-Maximization (EM) algorithm [101], have been developed to provide probabilistic estimates of missing values. However, these approaches still struggle with high-dimensional, heterogeneous data typical of AMR studies.

Machine learning has opened new opportunities for addressing this issue. Algorithms such as k-nearest neighbors (KNN) [10], random forests (RF) [11], and deep learning architectures like autoencoders and variational autoencoders (VAEs) [38] have demonstrated improved performance in imputing missing values by recognizing complex, non-linear dependencies among features. Generative frameworks such as Generative Adversarial Imputation Nets (GAIN) extend this capability by reconstructing missing data while preserving the overall distribution [121].

Although research specifically targeting missing data in AMR is still limited, related work in biomedical contexts provides a valuable foundation. Studies on electronic health records (EHRs) and genomic data analysis highlight the advantages of ML-based imputation methods in preserving correlations and improving predictive accuracy [59, 13]. Recent findings further suggest that hybrid models, which combine multiple imputation techniques or integrate statistical and ML approaches, may offer robust solutions under

different missingness scenarios [98, 103].

The application of ML to missing data extends beyond classical imputation. Ensemble-based strategies, such as Missing Imputation with Autoencoder and Ensemble Clustering (MIAEC) [120] or perturbation-driven clustering models [113], have been shown to handle uncertainty and noise effectively, particularly in large-scale, high-dimensional biological datasets. Deep learning approaches, including recurrent neural networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, have demonstrated strong robustness in dealing with structured and longitudinal datasets characterized by non-random missingness [122]. Semi-supervised frameworks, such as Iterative Robust Semi-Supervised Imputation (IRSSI), further extend these capabilities by leveraging both labeled and unlabeled data to improve accuracy under high missingness rates [34].

These methodological advances represent promising directions for AMR research, where incomplete and heterogeneous datasets are common. Nonetheless, challenges remain. Computational efficiency, model interpretability, and the risk of introducing artificial patterns during imputation are significant concerns that must be carefully managed to ensure the reliability of downstream analyses.

2.3 Limitations of ML in AMR Research

These advances illustrate the growing importance of ML in addressing the multifaceted challenge of AMR. On one side, predictive algorithms accelerate the identification of resistant strains from genomic data, offering a viable complement to traditional laboratory diagnostics. On the other hand, ML-based imputation strategies enhance the usability of incomplete datasets, a critical step for enabling large-scale surveillance and integrative modeling.

Despite these promising developments, several practical and methodological limitations still hinder the full integration of ML into clinical workflows. First, AMR datasets are frequently affected by severe data imbalance, where resistant and susceptible classes are unequally represented. This imbalance can lead to biased learning and unreliable predictions, particularly for underrepresented resistance phenotypes. Second, there is no universally optimal ML algorithm for AMR prediction. Model performance depends heavily

on dataset size, feature composition, and biological context, meaning that algorithm selection typically requires empirical comparison and extensive tuning. Finally, although ML models can generate predictions in seconds once trained, the upstream acquisition of genomic data through sequencing and bioinformatic analysis remains relatively slow and resource-intensive. This temporal bottleneck limits real-time clinical applicability, especially in urgent treatment settings.

While substantial evidence supports the potential of ML in AMR prediction, significant challenges remain, including the need for standardized benchmarks, interpretability of models, and clinical validation across diverse healthcare contexts. Nonetheless, the field is evolving rapidly, and ML-based frameworks are poised to play a central role in both the scientific understanding and the clinical management of antimicrobial resistance. Emerging solutions to mitigate current obstacles include the adoption of faster and cheaper sequencing technologies, the use of federated learning to integrate data from multiple institutions while preserving privacy, and the development of interpretable AI systems capable of providing transparent and clinically actionable predictions [100, 48, 60].

Chapter 3

AMR Dataset Description

In this chapter, we describe the datasets used in our studies on AMR prediction and missing-data handling. We first introduce binary genomic datasets focused on *Klebsiella pneumoniae*, including a proprietary clinical collection (from the Biometec departement of the University of Catania), a publicly available cohort [9], and a synthetic dataset generated to preserve key statistical properties of the real data. We then present a public continuous dataset on environmental antibiotic resistance genes (ARGs) [119].

3.1 Binary Data

In our study, we use genomic data to predict the antibiotic resistance of *Klebsiella pneumoniae* bacterium to different antimicrobial agents and to demonstrate the effectiveness of techniques for binary data reconstruction, in the context of AMR datasets.

In AMR prediction, the use of genomic data is fundamental. Genomic sequences of bacteria, in fact, are essential for understanding the molecular mechanisms underlying their resistance to antibiotics. In particular, genomic data may include information about the DNA, genes involved in resistance, and genetic mutations. The data encompass details regarding the presence or absence of specific resistance determinants. Since our prediction work focuses on *Klebsiella pneumoniae*, we pay particular attention to genes that provide resistance to β-lactam antibiotics, such as carbapenemases and extended-spectrum β-lactamases. This is essential in the clinical domain due to the alarming rise in resistance, particularly in nosocomial settings,

to several commonly used antibiotics [76]. The quest for genes responsible for these resistances involves the use of whole genome sequencing through Next Generation Sequencing (NGS) [92].

Here we consider two real datasets comprising the genomic data of *Klebsiella pneumoniae* strains, and a synthetic dataset, created from the characteristics of real data. As for real datasets, the first contains the data collected by a research group at the Department of Biomedical and Biotechnological Science (Biometec). For simplicity, we will refer to this dataset as the Biometec dataset. This dataset was used only for the AMR prediction study, due to its small size. The second is a public dataset, analyzed in [9]. This dataset was used both for AMR prediction and to apply missing data reconstruction techniques. Furthermore, the synthetic dataset, used for missing data management, was constructed based on the characteristics of this public dataset.

3.1.1 Biometec Data

The Biometec dataset contains the genomic data of 57 strains of *Klebsiella pneumoniae*. For each strain, the dataset has information on the resistance/-susceptibility to 15 antimicrobial agents and the presence/absence of 34 resistance genes, divided into 9 categories as follows: β-lactamase, quinolones, aminoglycosides, fosfomycin, sulfonamide, phenicols, macrolides, tetracicline and others, as illustrated in Table 3.1. For the majority of the strains (50 out of 57), the data also contains information on virulence genes. Specifically, the dataset includes information on the presence or absence of 78 virulence genes. The strains have different sources; in particular, we have considered strains taken from blood, urine, respiratory tract, rectal swabs, and one strain from a burn swab.

The samples included in the study were part of a strain collection of the laboratory of Microbiology at Biometec (University of Catania). The strains were collected in different hospitals located in South Italy during May 2020 and July 2023. For each patient, a single strain was included in the dataset. Identification and antimicrobial susceptibility were previously performed by the VITEK 2® system (bioMerieux, Marcy l'Etoile, France) at the hospitals and re-confirmed by standard methods (EUCAST, 2024). VITEK 2® is a fully automated system that performs bacterial identification and antibi-

Type	Genes
β-lactamase	$bla_{\text{SHV-28}}$, $bla_{\text{SHV-106}}$, $bla_{\text{SHV-187}}$, $bla_{\text{SHV-205}}$, $bla_{\text{SHV-212}}$, $bla_{\text{KPC-3}}$, $bla_{\text{KPC-31}}$, $bla_{\text{KPC-34}}$, $bla_{\text{OXA-1}}$, $bla_{\text{OXA-9}}$, $bla_{\text{TEM-181}}$, $bla_{\text{CTX-M-15}}$
Quinolones	*qnrB17*
Aminoglycosides	*aph(3")-Ib*, *aph(3")-Ib10*, *aph(6)-Id*, *aac(3)-IIe*, *aac(6')-Ib-cr6*, *aac(6')-Ib10*, *aadA2*, *ant(2")-Ia*, *ant(3")-IIa*, *armA*
Fosfomycin	*fosA6*
Sulfonamide	*sul1, sul2, dfrA12, dfrA14, dfrA17*
Phenicols	*catB3, catI*
Macrolides	*mphE*
Tetraciclin	*tetA*
Others	*qacEdelta1*

Table 3.1: List of genes included in the Biometec dataset divided into nine categories: β-lactamase, quinolones, aminoglycosides, fosfomycin, sulfonamide, phenicols, macrolides, tetracicline and 'others'.

otic susceptibility testing. It uses Advanced Colorimetry™, an identification technology that enables identification of routine clinical isolates. Advanced Colorimetry provides high discrimination between species and low rate of multiple choice and misidentified species. Breakpoints of antibiotics for the interpretative criteria for clinical isolates were used according to the EUCAST v 14.0 (EUCAST, 2024).

The main features of the Biometec dataset are summarized in Table 3.2 where we have reported the resistant and susceptible strains for the antimicrobial agents considered in our AMR prediction study.

Antimicrobial agent	**Resistant strains**	**Susceptible strains**	**Type**
AMC	57	0	*β-lactamase*
FEP	57	0	
CAZ	57	0	
IMI	41	16	
AZT	55	2	
CZA	43	14	
MEM	44	13	
MEM/VAB	16	41	
CIP	57	0	*Fluorochinolone*
CN	54	3	*Aminoglycoside*
AK	45	12	
COL	13	44	*Colistin*
FOS	40	17	*Fosfomycin*
SXT	39	18	*Sulfonamide*

Table 3.2: Summary of the main features of the Biometec dataset. The following antimicrobial agents included in the dataset have been considered in our study: *amoxicillin clavulanic acid* (AMC), *cefepime* (FEP), *ceftazidime* (CAZ), *imipenem* (IMI), *aztreonam* (AZT), *ceftazidime/avibactam* (CZA), *meropenem* (MEM), *meropenem/vaborbactam* (MEM/VAB), *ciprofloxacin* (CIP), *gentamicin* (CN), *amikacin* (AK), *colistin* (COL), *fosmomycin* (FOS), *trimethoprim-sulfamethoxazole* (SXT). The total number of strains is 57, divided between resistant and susceptible.

Type	Genes
β-lactamase	$bla_{\text{SHV-1}}$, $bla_{\text{TEM-1}}$, $bla_{\text{CTX-M1}}$, $bla_{\text{OXA-48}}$
Quinolones	*qnrB*
Aminoglycosides	*aac(6')-Ib-cr, aadb*
Sulfonamide	*sul1, sul2*

Table 3.3: List of resistance genes included in the public dataset divided in four categories: β-lactamase, quinolones, aminoglycosides and sulfonamide.

3.1.2 Public Data

The second binary dataset (i.e. the public one) contains the genomic data of 127 *Klebsiella pneumoniae* strains taken from four different Catalonian hospitals collected over six months in 2016. Also in this case, the strains are divided into three different categories, based on the origin of the bacterium: strains taken from blood, urine, and the respiratory tract. The dataset includes the following information: resistance/susceptibility of each strain to 12 antimicrobial agents; presence/absence of 9 resistance genes for each strain, and presence/absence of 16 virulence genes for each strain.

In the study regarding the prediction of AMR, we mainly concentrated on data pertaining to resistance genes. Specifically, the nine resistance genes appearing in the dataset are reported in Table 3.3. In our analysis of missing data, we will instead consider data relating to the presence or absence of both virulence and resistance genes.

The main features of public datasets are summarized in Table 3.4, where we have reported the resistant and susceptible strains for the antimicrobial agents considered in our AMR prediction study.

3.1.3 Synthetic Data

The third binary dataset is a synthetic one. We used this dataset in our missing data reconstruction work, and in particular, we created this dataset to test the effectiveness of imputation methods on a larger dataset than the

Antimicrobial agent	**Resistant strains**	**Susceptible strains**	**Type**
AMC	50	77	*β-lactamase*
FEP	30	97	
CAZ	47	80	
IMI	14	113	
AZT	37	90	
CIP	52	75	*Fluorochinolone*
CN	21	106	*Aminoglycoside*
COL	2	125	*Colistin*
FOS	18	109	*Fosfomycin*
SXT	44	83	*Sulfonamide*

Table 3.4: Summary of the main features of the public dataset. The following antimicrobial agents included in the dataset have been considered in our study: *amoxicillin clavulanic acid* (AMC), *cefepime* (FEP), *ceftazidime* (CAZ), *imipenem* (IMI), *aztreonam* (AZT), *ciprofloxacin* (CIP), *gentamicin* (CN), *colistin* (COL), *fosfomycin* (FOS), *trimethoprim-sultamethoxazole* (SXT). The total number of strains is 127, divided between resistant and susceptible.

real datasets we have available.

To create the synthetic dataset, we start from the real public dataset and generate new data with the same degree of correlation.In particular, to calculate correlations, we used the Pearson correlation coefficient ρ, defined as:

$$\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\sigma_A}\right)\left(\frac{B_i-\mu_B}{\sigma_B}\right) \tag{3.1}$$

where μ_A and σ_A are the mean and standard deviation of the variable A, respectively, and μ_B and σ_B are the mean and standard deviation of B. We can also define the correlation coefficient in terms of the covariance of A and B:

$$\rho(A,B) = \frac{cov(A,B)}{\sigma_A\sigma_B}. \tag{3.2}$$

Once the correlations were calculated, a new synthetic dataset was developed using Python's Copulas library. Using this library, it is possible to use copulas to learn the distribution of the given numerical data, allowing to generate synthetic samples that retain the same statistical properties as the original data. The first step in creating the synthetic dataset consists of removing the columns with zero variance, as they do not provide useful information and can interfere with the synthetic data generation process. A multivariate Gaussian copula is then trained (using the `Gaussian Multivariate` class from `copulas. multivariate` module of Python), particularly suitable for effectively representing the dependence between the variables. Using the trained copula, new synthetic samples are generated, which are then binarized by means of a threshold. The threshold is calculated as the average of the values of the column of the original dataset corresponding to the new generated one, in order to maintain consistency with the real data. Finally, the previously removed columns with zero variance, containing values consistent with the reference dataset, are reinserted. In particular, we generated a new dataset with 1000 samples and the same number of features as the public one.

3.1.4 Comparability and experimental implications of binary datasets

The two *Klebsiella pneumoniae* datasets differ in sampling frame (South Italy, May 2020–July 2023, N=57 vs. Catalonia, 6 months in 2016, N=127),

gene panels (Biometec: 34 resistance genes in 9 categories plus 78 virulence genes available for 50/57 strains; Public: 9 resistance genes in 4 categories plus 16 virulence genes), and antibiotic panels. The two datasets have ten common antimicrobial agents, which allow us to compare the results obtained from the work of predicting AMR. In particular, the common antimicrobial agents are the following: *Amoxicillin/clavulanate* (AMC), *Cefepime* (FEP), *Ceftazidime* (CAZ), *Imipenem* (IMI), *Aztreonam* (AZT), *Ciprofloxacin* (CIP), *Gentamicin* (CN), *Colistin* (COL), *Fosfomycin* (FOS), *Trimethoprim/sulfamethoxazole* (SXT). However, given their clinical importance, in our study on the prediction of AMR, we have also considered four other antimicrobial agents that are only present in the Biometec dataset. These are: *Ceftazidime/Avibactam* (CZA), *Meropenem* (MEM), *Meropenem/Vaborbactam* (MEM/VAB) and *Amikacin* (AK). In fact, *Ceftazidime*, a 3rd generation cephalosporin, is an important antibiotic molecule used in clinical practice. This antibiotic is used in combination with *Avibactam*, an inhibitor of class A, C, and some class D β-lactamase. The antibacterial spectrum of *Ceftazidime-Avibactam* covers >99% of enterobacteria, including strains carrying extended-spectrum β-lactamase [77], such as *Klebsiella pneumoniae*, under examination in this study. The other three antibiotics play a prominent role in clinical practice, in particular *Meropenem* (MEM), is a β-lactam antibiotic belonging to broad-spectrum carbapenems, only and in combination with the β-lactamase inhibitor *Vaborbactam* (MEM/VAB) and *Amikacin* (AK), a semi-syntetic aminoglycoside antibiotic used for most resistant Gram-negative bacteria. Due to their activity against multidrug resistant (MDR) pathogens, all aforementioned antibiotics should be considered as some of the most important molecules in case of infections supported by MDR pathogens.

The synthetic dataset, as previously mentioned, was used only for the work of reconstructing the missing data, along with the public dataset. In this case, since the synthetic dataset is directly derived from the real one, the two datasets are comparable, allowing us to see how the results are affected by the size of the dataset under consideration.

3.2 Continuous Data

In the study of reconstruction of missing data, ML techniques were applied not only to binary data but also to continuous data. For this purpose, we employed a public dataset [119] describing the abundance of environmental antibiotic resistance genes (ARGs). The dataset represents one of the most comprehensive resources currently available for investigating antimicrobial resistance in natural ecosystems. It was compiled from 18 provinces across China between 2013 and 2020 and encompasses five major habitats (soil (SL), water (WT), sediment (SD), atmospheric particulate matter (PM), and dust (DT)), covering a total of 653 sampling sites and 1403 bacterial strains.

The study provides details on the spatiotemporal distribution of 290 ARG subtypes across 291,870 samples, along with the abundance of 30 mobile genetic elements (MGEs), including transposases, plasmids, insertion sequences, and integrases. Sampling and DNA extraction were standardized to ensure reliability and comparability across sites. ARG abundance was quantified through High-Throughput Quantitative Polymerase Chain Reaction (HT-qPCR) using the SmartChip Real-Time PCR system, which employed 414 primer pairs targeting ARGs, MGEs, and the 16S rRNA gene. Relative gene abundance was determined as the ratio between ARG copies and 16S rRNA copies, while absolute abundance was calculated by multiplying the absolute copy number of 16S rRNA by the relative abundance of each ARG.

Each sample in the dataset includes categorical or discrete features, such as gene subtype, resistance mechanism, and acquisition province, and continuous features, including relative and absolute abundance. The ARGs identified confer resistance to 15 major antibiotic classes, such as β-lactams, aminoglycosides, and fluoroquinolones, and were further grouped according to six resistance mechanisms, including enzymatic antibiotic deactivation and reduced membrane permeability. Additionally, ARGs were categorized into four risk ranks according to their potential threat to human health: Rank I, ARGs associated with ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter*); Rank II, ARGs not

yet detected in pathogens; Rank III, ARGs associated with humans but not with MGEs; and Rank IV, ARGs not associated with humans.

Starting from this original dataset, we extracted a derived dataset containing information on the relative abundance of genes for each of the 1403 bacterial strains examined. In particular, 5 sub-datasets were then obtained, each corresponding to a different habitat and containing continuous values relating to the relative abundances of each gene within the bacterial strain. These datasets were obtained by performing data segmentation and data filtering operations on the original dataset, such as feature removal, sample removal, and dataset transposition. The characteristics of the 5 datasets are reported in Tab 3.5.

Habitat	**Samples**	**Features**
SL	691	50
WT	227	102
SD	70	68
PM	28	28
DT	117	32

Table 3.5: Number of samples (bacterial strains) and characteristics (relative abundance of genes present) in the 5 ARG datasets with continuous values of gene abundances for different habitats: soil (SL), water (WT), sediment (SD), atmospheric particulate matter (PM), and dust (DT).

Chapter 4

Machine Learning models

In this chapter, we introduce the fundamental principles of ML and its application to the analysis of biological and biomedical data. We begin with a general overview of ML, illustrating its main paradigms and the typical workflow that underlies the construction, training, and validation of predictive models. Subsequently, we focus on two specific applications relevant to this study: the use of ML techniques for *classification*, aimed at assigning samples to predefined categories, and for *data imputation*, intended to estimate missing values and reconstruct incomplete datasets. For each task, the most representative algorithms adopted in this work are presented, describing their theoretical foundations, main assumptions, and typical advantages and limitations.

4.1 Prediction and Classification in Machine Learning

Depending on the availability of labeled data and on the learning objective, ML algorithms can be grouped into different paradigms: supervised, unsupervised, semi-supervised, and reinforcement learning [1, 118]. Supervised and semi-supervised approaches are those most relevant to this study, as they rely on partially or fully labeled datasets to infer mappings between input features and output variables. Within supervised and semi-supervised learning frameworks, machine learning tasks are commonly divided into two principal categories: *classification* and *prediction* [123, 1, 118]. Although

both involve learning from labeled data, they differ in the nature of the target variable and in their intended purpose.

In *classification*, the goal is to assign input samples to one or more predefined discrete categories. The target variable is therefore categorical, and the algorithm learns a decision boundary that best separates the classes in the feature space. In the AMR context, typical examples include the identification of bacterial species based on genomic features or the categorization of samples as resistant or susceptible to a specific antibiotic. The performance of classification models is usually evaluated through metrics such as accuracy, precision, recall, F1-score, or the area under the receiver operating characteristic curve (AUC).

Conversely, *prediction* tasks, also referred to as *regression* when the output is continuous, aim to estimate a numerical value rather than a discrete label. These models are used to predict quantitative traits, such as (in the case of AMR datasets) the expression level of a gene, the minimum inhibitory concentration (MIC) of an antibiotic, or a missing measurement within a dataset. When prediction is applied to the estimation of missing data points, as in the case discussed in this thesis, the process is often termed *imputation*. In this context, ML algorithms learn the statistical relationships among observed variables to infer plausible values for unobserved ones, thus improving data completeness and preserving the multivariate structure of the dataset.

While classification focuses on discrete decision-making and pattern discrimination, prediction and imputation emphasize quantitative estimation and reconstruction of underlying relationships. Both paradigms rely on the same fundamental ML principles (training on representative data, generalization to unseen cases, and model evaluation) but differ in their objectives and in the choice of algorithms and performance metrics appropriate to the problem.

4.2 Machine Learning methods for classification

Machine learning algorithms for classification are designed to assign input data to one or more predefined categories based on a set of descriptive features. In supervised classification, the model learns from labeled examples

$(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ represents the feature vector and y_i the corresponding class label. The goal is to approximate the underlying function $f : \mathbf{x} \to y$ that maps input features to output categories, minimizing classification error on unseen data [50]. Classification models differ significantly in their underlying assumptions, mathematical formulation, and interpretability. Broadly, they can be grouped into four main families [117, 88]:

- *Probabilistic models:* derive class membership probabilities from the statistical distribution of features within each class, typically based on Bayes' theorem (e.g., Naive Bayes classifiers).
- *Linear models:* define decision boundaries as linear combinations of features and optimize parameters to separate classes according to a discriminant function (e.g., Logistic Regression).
- *Instance-based models:* make predictions based on the similarity between new samples and stored examples, without learning explicit global parameters (e.g., k-Nearest Neighbors and Radius Neighbors).
- *Ensemble models:* combine multiple base estimators to enhance predictive accuracy and robustness, either through parallel aggregation (Bagging) or sequential boosting (Gradient Boosting).

Each family offers specific advantages depending on the data structure, dimensionality, and class separability. Probabilistic and linear models tend to be more interpretable and computationally efficient, while instance-based and ensemble approaches are generally more flexible and capable of modeling non-linear relationships.

In this thesis, we analyze six machine learning models: Gaussian Naive Bayes, Logistic Regression, k-Nearest Neighbors, Radius Neighbors, Gradient Boosting, and Bagging Classifier [123]. A theoretical overview of each method is reported below.

4.2.1 Gaussian Naive Bayes

Naive Bayes classifiers are probabilistic models based on Bayes' theorem, which describes the posterior probability of a class C_k given the observed

features $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ as:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})} \tag{4.1}$$

The defining assumption of the Naive Bayes approach is the conditional independence of the features given the class, i.e.:

$$P(\mathbf{x}|C_k) = \prod_i P(x_i|C_k) \tag{4.2}$$

Despite its simplicity, this assumption often performs surprisingly well in high-dimensional problems. In the Gaussian Naive Bayes variant, each feature is assumed to follow a Gaussian (normal) distribution within each class, allowing analytical computation of the likelihood. The model is computationally efficient, interpretable, and particularly effective when features are approximately normally distributed and class boundaries are relatively simple.

4.2.2 Logistic Regression

Logistic regression is a parametric model that estimates the probability of a categorical outcome as a logistic (sigmoid) function of a linear combination of input features. For a binary classification task, the model can be expressed as:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}} \tag{4.3}$$

where $\boldsymbol{\beta}$ represents the vector of learned coefficients. The decision boundary corresponds to the hyperplane where $P(y = 1|\mathbf{x}) = 0.5$. Although conceptually simple, logistic regression remains a powerful baseline due to its interpretability, statistical robustness, and well-understood optimization behavior. Extensions such as multinomial logistic regression generalize the model to multi-class problems by estimating probabilities across several mutually exclusive outcomes.

4.2.3 k-Nearest Neighbors

The k-Nearest Neighbors (KNN) algorithm is a non-parametric, instance-based method that classifies a new sample based on the majority vote of its

k closest samples in the training set, according to a chosen distance metric (typically Euclidean). Unlike parametric models, KNN makes no explicit assumptions about the underlying data distribution; instead, it relies entirely on the local structure of the feature space. Its performance is highly dependent on the choice of distance metric, the value of k, and the scaling of features. The algorithm is simple to implement and effective when the decision boundaries are irregular, but it can become computationally expensive for large datasets, as the classification of each new instance requires computing distances to all training samples.

4.2.4 Radius Neighbors Classifier

The Radius Neighbors Classifier is a variant of instance-based learning similar to KNN. Rather than using a fixed number of neighbors, it classifies each sample based on all training instances lying within a specified distance (radius) r. This approach allows the local density of the data to influence the classification: dense regions with many nearby points contribute more evidence than sparse regions. In cases where no neighbors fall within the defined radius, the model may abstain from classification or apply alternative strategies. The method is well-suited to datasets with non-uniform sampling density, where a fixed k might be inappropriate across different regions of the feature space.

4.2.5 Bagging Classifier

Bootstrap Aggregating, or *Bagging*, is an ensemble technique designed to reduce the variance of unstable models by combining multiple estimators trained on random bootstrap subsets of the training data. Each base model is trained independently, and the final prediction is obtained by averaging (for regression) or voting (for classification) across all models. This approach increases robustness and generalization, particularly when the base estimators are sensitive to small variations in the data (e.g., decision trees or support vector machines). Bagging typically improves predictive accuracy at the cost of increased computational complexity, as it requires training multiple models.

4.2.6 Gradient Boosting Classifier

Gradient Boosting is a sequential ensemble method that aims to reduce both bias and variance by combining multiple weak learners, usually shallow decision trees, into a single strong predictor. Each new model is trained to minimize the residual error of the previous ensemble, effectively performing a form of gradient descent in function space. The overall model can be expressed as:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}) \tag{4.4}$$

where h_m is the weak learner fitted to the residuals of the current model and γ_m is a scaling factor controlling its contribution. Gradient Boosting is particularly powerful for complex, non-linear classification problems, as it adaptively focuses on difficult-to-classify samples. However, it requires careful tuning to prevent overfitting, and its sequential nature leads to higher computational cost compared to parallel ensemble methods such as Bagging.

4.3 Machine Learning methods for missing data imputation

Missing data imputation encompasses a collection of statistical and computational techniques aimed at estimating and replacing missing values in a dataset with plausible alternatives. The objective of these methods is to reconstruct the integrity of incomplete datasets by exploiting the relationships and dependencies among observed variables, thus reducing information loss and mitigating potential biases introduced by missingness [33]. Imputation allows for the preservation of the entire data set, unlike listwise deletion or case-wise removal approaches, which can significantly reduce sample size and statistical power.

From a methodological point of view, imputation approaches can be classified along multiple dimensions. Depending on the number of variables considered in the estimation, it is possible to distinguish between *univariate* methods, which impute missing values using information from the same variable only, and *multivariate* methods, which leverage correlations among several variables. Another important distinction is between *single* and *multiple* imputation. Single imputation fills each missing entry once, providing a

complete dataset but underestimating uncertainty. Multiple imputation, by contrast, generates several imputed datasets under slightly different assumptions and combines them to produce a more robust and unbiased estimate of the missing values [57, 30].

A further categorization can be drawn between *classical statistical* and *machine learning-based* imputation techniques. Classical methods, such as mean or regression imputation, rely on simple distributional assumptions and are computationally efficient but limited in their ability to capture non-linear or high-order relationships among variables [51, 12]. ML-based techniques, on the other hand, leverage flexible, data-driven models capable of identifying complex multivariate dependencies. They often achieve superior accuracy and generalization, particularly in heterogeneous biomedical datasets where missingness mechanisms are not purely random [61, 71].

Broadly, the main families of imputation methods can be summarized as follows:

- *Statistical (classical) methods:* include simple techniques such as mean, median, or mode substitution, regression-based imputation, and Last Observation Carried Forward (LOCF). These are computationally efficient but may introduce bias or reduce data variability.
- *Machine learning-based methods:* exploit predictive models to estimate missing values using available features. Examples include k-Nearest Neighbors (KNN), Random Forest (RF), Multiple Imputation by Chained Equations (MICE), and clustering-based methods such as k-Means Imputation (KMI). These methods capture non-linear relationships and preserve multivariate structures, albeit at higher computational cost.

The following sections describe the theoretical foundations and main characteristics of the four ML-based methods used in this study.

4.3.1 k-Nearest Neighbors Imputation

The KNN algorithm imputes missing values by exploiting the similarity between samples. For each instance containing missing entries, the algorithm identifies the k most similar instances in the dataset, based on a chosen

distance metric such as the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.5}$$

The missing value is then estimated as a weighted average of the corresponding feature values from these neighbors, with weights typically inversely proportional to distance. This approach assumes that similar samples exhibit similar feature values, thereby preserving the local structure of the data.

KNN imputation is simple, non-parametric, and effective for datasets with low to moderate dimensionality. However, it can be computationally demanding for large datasets and is sensitive to the choice of k and the distance metric. A small k captures local variability but may overfit noise, whereas a large k yields smoother imputations at the cost of reduced sensitivity to local structure.

4.3.2 Random Forest Imputation

RF imputation is based on an ensemble of decision trees, each trained on bootstrap samples of the data. In this approach, missing values are estimated by predicting them with a RF model that uses the observed features as predictors. Each decision tree partitions the data recursively according to feature thresholds, generating multiple independent estimates that are then aggregated, typically by averaging (for continuous variables) or by majority vote (for categorical variables). The ensemble nature of RF allows it to capture complex, non-linear relationships between variables while reducing the risk of overfitting compared to single-tree models.

RF imputation is particularly effective when variables exhibit strong multivariate dependencies and heterogeneous data distributions. Its main advantages include robustness to outliers, the ability to handle mixed-type variables, and the preservation of variable interactions. However, it requires substantial computational resources and can introduce bias if the number of trees or their depth is not properly controlled.

4.3.3 Multiple Imputation by Chained Equations

MICE is an iterative, model-based approach that performs multiple rounds of imputation to approximate the joint distribution of the data [56, 102]. Each variable with missing values is modeled conditionally on all other variables through a sequence of regression models, forming a chain of equations. At each iteration, the missing values for one variable are imputed based on the most recent imputations of the others, and the process is repeated cyclically until convergence.

MICE can employ different regression models depending on the type of variable, for instance, linear regression for continuous data or logistic regression for categorical data. Its iterative nature allows it to capture complex inter-variable dependencies and quantify uncertainty by generating multiple imputed datasets. However, MICE assumes that the data are Missing Completely at Random (MCAR) or Missing at Random (MAR), and its computational cost can be high for large datasets or highly correlated variables.

4.3.4 k-Means Cluster Imputation

The KMI method relies on partitioning the dataset into clusters of similar observations using the k-Means clustering algorithm. Once the data are grouped, missing values are imputed using the statistics (typically the mean) of the corresponding variable within each cluster. Formally, for a dataset partitioned into n_k clusters C_i with centroids μ_i, the within-cluster sum of squared errors (SSE) is minimized:

$$\text{SSE} = \sum_{i=1}^{n_k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4.6}$$

The imputed value for a missing feature is thus derived from the mean of that feature among the non-missing entries within the same cluster.

The effectiveness of KMI depends critically on the selection of the number of clusters n_k and the distance metric used to define similarity. An inappropriate choice may result in over- or under-partitioning of the data and, consequently, poor imputations. Despite this, KMI can efficiently capture local group structures in high-dimensional data and is particularly suitable when the dataset contains well-defined subpopulations.

Together, these imputation algorithms provide complementary strategies for handling missing data. Instance-based and clustering-based approaches emphasize local data structure, ensemble methods capture complex non-linear relationships, and iterative model-based algorithms offer statistically grounded uncertainty estimation. In combination, they represent a comprehensive toolkit for addressing the pervasive issue of incomplete data in biomedical and genomic analyses.

Chapter 5

Performance Evaluation Metrics

Depending on the nature of the task (classification or regression) different metrics are employed to quantify the agreement between the predicted and the true values. In this work, performance assessment was conducted for both classification models (used for AMR prediction) and regression or reconstruction models (used for missing data imputation). In what follows, we describe the evaluation metrics adopted in both contexts, distinguishing between those suitable for binary classification and those used for continuous-valued data reconstruction.

5.1 Metrics for Binary Classification Tasks

Binary classification models, such as those applied to AMR prediction or to binary gene presence/absence imputation, produce categorical outcomes that can be summarized through a confusion matrix. Let TP (True Positives) denote the number of correctly predicted positive samples, TN (True Negatives) the correctly predicted negatives, FP (False Positives) the incorrectly predicted positives, and FN (False Negatives) the incorrectly predicted negatives. From these quantities, several performance indicators can be derived, as reported in the following sections.

5.1.1 Accuracy

Accuracy can be defined as the ratio of the number of correct predictions to the total number of predictions:

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of predictions}} \tag{5.1}$$

Equivalently, the accuracy through the outcome of the model prediction can also be computed as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5.2}$$

This metric provides a general measure of overall correctness. However, it can be misleading in the presence of class imbalance, since it does not account for the relative distribution of false positives and false negatives.

5.1.2 Precision and Recall

To address this limitation, additional measures are employed to capture the model's behaviour with respect to specific error types. *Precision* quantifies the proportion of true positives among all predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{5.3}$$

while *Recall* (also known as Sensitivity or True Positive Rate) measures the proportion of true positives that were correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{5.4}$$

In biomedical contexts, such as AMR prediction, recall is particularly relevant because false negatives (i.e., resistant bacteria incorrectly predicted as susceptible) may have critical clinical implications.

5.1.3 F1-Score

To provide a single indicator that balances precision and recall, the *F1-score* is computed as their harmonic mean:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.5}$$

The F1-score is especially informative when dealing with imbalanced data or when both types of errors (FP and FN) carry comparable importance.

5.1.4 Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC)

Another widely used approach to assess binary classifiers is the analysis of the *Receiver Operating Characteristic* (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The *Area Under the Curve* (AUC) provides a threshold-independent measure of the model's discriminative ability:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{5.6}$$

The AUC value ranges between 0 and 1, where 0.5 indicates random performance, and 1.0 represents perfect discrimination between classes. A higher AUC denotes a better capacity of the model to correctly distinguish between positive and negative samples.

5.2 Metrics for Continuous-Valued Imputation Tasks

When the task involves reconstructing continuous values, such as the estimation of missing entries in a dataset, the evaluation must rely on error-based metrics comparing the predicted and the true numerical values. Given y_i as the true value and $\hat{y}_i$ as the predicted (or imputed) value, the error metrics reported below can be used.

Among the most commonly used error metrics are the *Root Mean Square Error* (RMSE), the *Mean Absolute Error* (MAE), and the *Mean Relative Error* (MRE), which quantify the deviation between observed and predicted values in different ways.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.8}$$

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{5.9}$$

RMSE penalizes large errors more heavily due to the quadratic term, making it sensitive to outliers, whereas MAE provides a more robust measure of the average magnitude of the errors. The MRE, on the other hand, expresses the average error relative to the true value, allowing an interpretation in percentage terms. However, this metric can become unstable or undefined when the true values approach zero.

While these metrics are informative, their absolute values depend on the data scale, making cross-dataset comparisons difficult. Since raw error magnitudes can depend on the scale of the data, it is often preferable to use normalized versions of the previous metrics, such as the *Normalized Root Mean Square Error* (NRMSE) and the *Normalized Mean Absolute Error* (NMAE):

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}} \tag{5.10}$$

$$\text{NMAE} = \frac{1}{n} \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{y_{\max} - y_{\min}} \tag{5.11}$$

Normalizing with respect to the data range makes these error measures dimensionless and thus comparable across datasets characterized by different scales or units. The NMAE was preferred over other relative error measures, such as the MRE, due to its greater robustness and interpretability, especially in cases where true values may approach zero, potentially leading to unstable or undefined MRE values.

Chapter 6

AMR prediction

This chapter describes the methodological setting and the main results obtained in the prediction of AMR in *Klebsiella pneumoniae* isolates, using ML approaches. The analysis starts from the genomic characterization of bacterial isolates and the generation of sequencing data, through DNA extraction and Next Generation Sequencing, to the bioinformatic processing and annotation of genomic features. The resulting datasets were subsequently preprocessed to ensure data quality and relevance, including correlation analysis and class balancing procedures. Different ML models were then designed, trained, and optimized to predict resistance phenotypes based on genomic information. Finally, we present and discuss the obtained results, highlighting the main findings, the performance of the proposed models, and the biological and methodological implications of the study.

For the AMR prediction analysis conducted in this chapter, only the two real binary datasets (Biometec dataset and public dataset) were used.

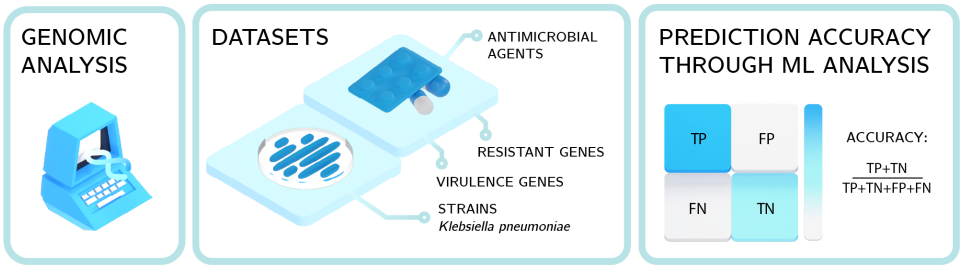


Figure 6.1: **Synthetic workflow for AMR prediction from genomic datasets.**

6.1 Genomic analysis

Genomic analysis of bacteria is a complex and detailed process to identify their genetic code. DNA sequencing was performed by means of a new technology called Next-Generation Sequencing (NGS) [92]. This technology allows large numbers of genes to be sequenced in short periods of time. The process begins with the extraction of DNA from the bacterium of interest, which can be done from bacterial cultures or samples. The extracted DNA is then split into small segments which are then treated with different techniques before moving to the sequencing phase. The sequencing phase occurs in parallel and is done using techniques based on different principles. In our case Illumina® (fluorescent sequencing, short reads) was used. The sequences generated by NGS are then analyzed to search for genes with specific characteristics, such as antibiotic resistance or virulence.

Genomic analysis was performed on strains collected at the laboratory of Medical Molecular Microbiology and Antibiotic Resistance (MMAR) at Biometec, according to the three steps that are discussed below.

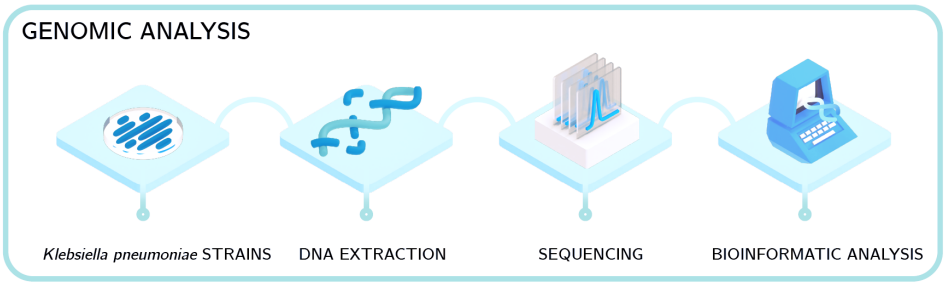


Figure 6.2: **Genomic analysis workflow.**

6.1.1 DNA extraction

The first step is DNA extraction from the sample of interest. DNA extraction was carried out following the manufacturer's instructions provided by QIAGEN QIAamp® DNA Mini Kit (Ref. 51304, QIAGEN, 40724 Hilden, Germany). DNA was quantified using both the Eppendorf BioPhotometer® D30 and the fluorimeter Qubit dsDNA BR Assay Kit to evaluate purity and quantity of the initial sample, respectively (Ref. 32850, Invitrogen, 92008 Carlsbad, CA, USA)[14].

6.1.2 Next Generation Sequencing (NGS)

A concentration of $100ng$ of each sample was used for NGS. This step was performed in the laboratory of Molecular Biology at the University of Catania on an Illumina® MiSeq platform according to the manufacturer's instructions provided in Watchmaker DNA Library Prep kit with Fragmentation–Watchmaker Genomics® (Ref. 7K0013-024, 5744 Central Avenue, Suite 100 Boulder, CO 80301, USA). Indexes were provided with Twist Universal Adapter System (16 Indexes, 16 Samples) (Ref. 101307, Twist Bioscience, HQ 681 Gateway Blvd, South San Francisco, CA 94080FAQ). Libraries were quantified and their quality was evaluated using both the fluorometric Qubit dsDNA HS Assay Kit (Ref. Q32851, Invitrogen, Carlsbad, CA 92008, USA) and the Agilent® High Sensitivity DNA Kit (Ref. 5067-4626). Denature and dilute libraries were performed following the "Denature and Dilute Libraries Guide" protocol provided by Illumina®, choosing 8,5 pM as the loading concentration. Finally, sequencing was performed using the MiSeq Reagent Kits v3 (Ref. 15043895, Illumina, Inc., 92122, San Diego, CA, USA). The Sample Sheet was created using the Local Run Manager v3 software, and following the instructions in the Local Run Manager v3 Software Guide provided by Illumina®.[14]

6.1.3 Bioinformatic analysis

Data were analyzed using the QIAGEN CLC Genomics Workbench software and following the user manual for the CLC Microbial Genomics Module 23.0.2, released on July 7, 2023 (QIAGEN, Aarhus, 8000 Denmark), to assign resistance gene, virulence and Multi-locus Sequence Type (MLST).

6.2 Preprocessing of the datasets

Before applying machine learning algorithms, it is essential to properly preprocess the data in order to ensure the robustness and reliability of the subsequent analysis. Preprocessing involves a series of operations designed to remove redundant or irrelevant information, mitigate biases, and improve data quality. In this section, we describe the procedures implemented prior to the model training phase, with particular attention to the correlation

analysis and the balancing of the dataset.

6.2.1 Correlation analysis

From the public database described in Section 3.1.2, we extracted a subset of data through a preliminary correlation analysis. The objective of this step was to identify which features (genes) are most relevant for predicting antimicrobial resistance and to exclude those not contributing to the classification task. In particular, virulence genes were discarded since they do not provide informative patterns for resistance prediction, whereas resistance genes are expected to be directly associated with the antimicrobial phenotype.

To this aim, we calculated the Pearson correlation coefficient [54] using MATLAB®. The correlation coefficient measures the linear dependence between two random variables. Assuming that each variable has N scalar observations, the Pearson correlation coefficient ρ is defined as in equations 3.1–3.2. The Pearson correlation coefficient ρ has been used to evaluate the correlation between the resistance to each antimicrobial agent and the presence of each resistance or virulence gene. Following [43], values of $|\rho| > 0.8$ were considered indicative of strong correlation, $0.4 < |\rho| < 0.8$ of moderate correlation, and $0 < |\rho| < 0.4$ of weak correlation. The analysis revealed no significant correlation between virulence genes and antimicrobial resistance, whereas a strong to moderate correlation was observed between resistance genes and specific antimicrobial agents (Fig. 6.3). Based on these findings, only resistance genes were retained for the subsequent analysis.

6.2.2 Data balancing

A common issue in applying ML methods to biological datasets is class imbalance, where one class (e.g., resistant strains) is overrepresented relative to the other (e.g., susceptible strains). Such imbalance can bias the model, leading it to favor the majority class during training and consequently reducing its predictive power for the minority class. Balancing the dataset mitigates this problem by ensuring that each class contributes equally to the learning process.

Among the available approaches, oversampling techniques replicate or

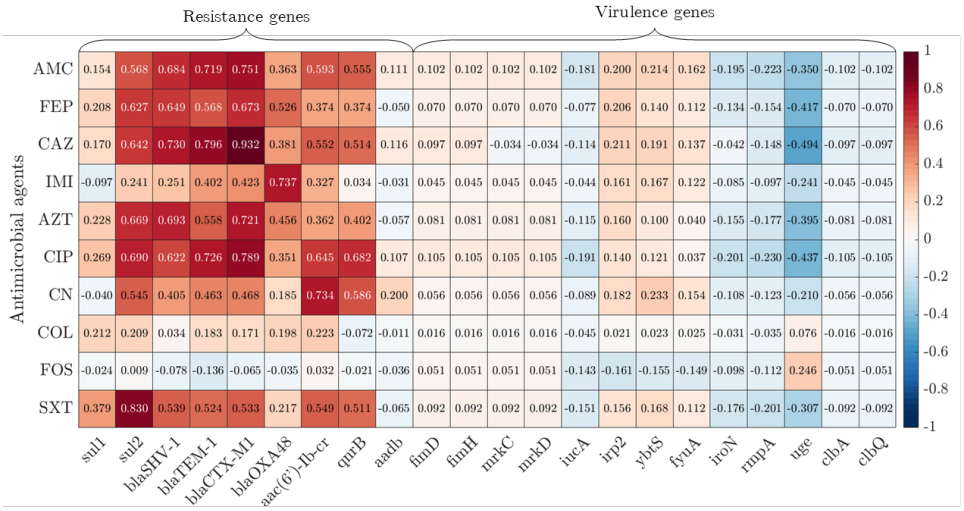


Figure 6.3: **Pearson correlation coefficient.** Values of the Pearson correlation coefficient ρ obtained for the public dataset. We find a strong correlation ($|\rho| > 0.8$) between resistance genes (*sul1*, *sul2*, $bla_{\text{SHV-1}}$, $bla_{\text{TEM-1}}$, $bla_{\text{CTX-M1}}$, $bla_{\text{OXA-48}}$, *aac(6')-Ib-cr*, *qnrB*, *aadb*) and resistance to antimicrobial agents. Virulence genes are less correlated with resistance to antimicrobial agents, or at most, the correlation is negative (the presence of the genes may imply non-resistance to antimicrobial agents). We also notice a low correlation between the genes and the antimicrobial agent FOS. Virulence genes ycfM, entB, and wabG are not reported because they are present in all strains (all 1), which means that $\sigma = 0$, such that ρ cannot be evaluated.

synthetically generate samples of the minority class to match the size of the majority class. However, simple replication may cause overfitting; therefore, data augmentation techniques that introduce variability among the generated samples are preferable.

In extreme cases, where all samples belong to a single class, the associated antibiotic was excluded from the analysis before model training. For the remaining antibiotics, we applied the Synthetic Minority Oversampling Technique (SMOTE) [20], which generates synthetic samples in the feature space by interpolating between existing minority class instances. This method effectively reduces class imbalance and improves model generalization without significantly increasing the risk of overfitting.

6.3 Machine learning models setting

As mentioned in Section 4.2, in this work, we use supervised learning for classification. In our research, the initial phase involves data preprocessing (6.2). As already mentioned, for balancing data we used the SMOTE. SMOTE is implemented in Python in the `imbalanced-learn` library, which can be used in conjunction with scikit-learn. SMOTE is helpful in particular because it balances the class distribution by creating synthetic samples for the minority class by interpolating between samples of the minority class that already exist.

Subsequently, we applied the ML methods described in Section 4.2. The input sequences used for the AMR classification with ML consist of data indicating the presence or absence of resistance genes in individual bacterial strains. For each strain, we have data to define whether a specific gene was present (denoted as 1) or absent (denoted as 0). As result of the training step, we get the resistance, indicated as 1, or susceptibility, indicated as 0, of the strains to each of the antimicrobial agents considered in our analysis.

The code used in our work is written in Python, in particular, we used the scikit-learn Python pakage. For Gaussian Naive Bayes we use the `GaussianNB` from the package `sklearn.naive_bayes` of Python, with default parameters. The second model used is Logistic regression. The logistic regression is implemented in Python in `LogisticRegression`, from `sklearn.linear_model`. We use it in "multinomial" `multi_class` and with a maximum number of iterations equal to 3000. Regarding the neighbors-based models, we used `sklearn.neighbor`, and in particular `KNeighborsClassifier`, that implements learning based on the `k` nearest neighbors of each query point (`k` integer value), and `RadiusNeighborsClassifier`, that implements learning based on the number of neighbors within a fixed radius `r` of each training point (`r` floating-point value). In both cases we test the models with different parameters and we choose the best solution with the `GridSearchCV` function, located in `sklearn.model_selection`. In particular, for the `KNeighborsClassifier` we set `n_neighbors` parameter equal to 2,6 or 10, while for `weight` and `algorithm` parameters we tested all the possible values (`distance` and `uniform` for the weight function used in prediction, and `auto`, `ball_tree`, `kd_tree` and `brute` for the algorithm

used to compute the nearest neighbors). For the other parameters we used default values. For `RadiusNeighborsClassifier` we set the `radius` parameter in a range of 9, 30 or 1, while for the `weight` and `algorithm` parameters we tested the same values as in the previous case. The other parameters are set by default. In order to compute the ensemble methods we used the `BaggingClassifier` and `GradientBoostingClassifier` from `sklearn.ensamble`. For `BaggingClassifier` we use SVC (from `sklearn.svm`) as base estimator. Then we set `n_estimator=10`, `max_saples=0.8` and `max_fearures=0.8`. As an example of the second family, we use the `GradientBoostingClassifier` where we set the number of boosting stages equal to `n_estimetors=75`, `validation_fraction=0.2`, `n_inter_no_change=5`, `tol=0.01` and `randm_state=0`.

6.4 Results

The first step of our analysis was to eliminate the antimicrobial agents associated with bacterial strains that are all resistant or susceptible to a specific antimicrobial agent. In the Biometec dataset (Table 3.2) we find that for four antimicrobial agents only resistant strains are available: AMC, FEP, CAZ and CIP. Since we cannot predict the resistance or susceptibility for these antimicrobial agents, they have not been considered in the other steps of our analysis.

We first consider the application of the ML methods to the original unbalanced datasets. Using the original datasets, we find that for some antimicrobial agents several of the methods of Section 4.2 fail to converge. In other cases, even if the methods converge, the accuracy values are on average around 10% less, with typically a large number of FP and FN. This confirms the need of balancing for our data.

For this reason, we decided to use a technique for data balancing, in particular SMOTE as discussed in Section 6.2. In this way, applying ML methods to data, we ensure that learning is possible also for the minority class, which also improves the values of precision, recall, and accuracy of the predictions.

Next, we illustrate the results for the balanced data. We anticipate that, comparing the results for all the ML methods of Section 4.2, the Gradient

Boosting classifier and k-nearest neighbors classifier are generally the two methods that better perform, i.e., the two which give the highest accuracy in the prediction.

We begin with the analysis carried out on the public dataset. The dataset, as described in Section 3.1.2, contains information on the susceptibility or resistance of 127 bacterial strains of *Klebsiella pneumoniae* to 12 different antimicrobial agents. However, as previously discussed, we consider the data only of the 10 antimicrobial agents that are present also in the Biometec dataset. For these agents we apply the six ML methods described in Section 4.2, and for each of these methods we evaluate the accuracy (5.2). The results are illustrated in Table 6.1 and in Fig. 6.4, where they are also compared with the same analysis carried out on a smaller number of strains. In more detail, to test the viability of the method as a function of the size of the analyzed sample, we have applied the same procedure used for the entire dataset to a smaller subset of data. Specifically, we focused the analysis on the 37 strains of the respiratory tract. For the respiratory tract, in fact, all antimicrobial agents, with the only exception of COL, for which there is a single resistant strain, have a sufficient number of resistant and sensitive strains.

Table 6.1 shows that, in the case of the analysis of the whole dataset, the values of accuracy are above 90%. It happens for all methods, with the Gradient Boosting classifier performing better than the other techniques for almost all antibacterial agents. The comparison (see also Fig. 6.4) with the accuracy values obtained on the smaller dataset of respiratory strains shows that no clear differences emerge when the whole dataset or a part of it is used for the analysis. We conclude that predicting antimicrobial resistance from genomic data is possible, even when a small number of strains is available, e.g. the subset of the 37 strains related to the respiratory tract bacteria.

Next, we discuss the analysis of the Biometec dataset, which, as described in Section 3.1.1, contains information on the resistance or susceptibility to different antimicrobial agents of 57 bacterial strains. The results are illustrated in Table 6.2. The highest accuracy values are obtained for k-nearest neighbors ML model, with all other models also displaying good performances (accuracy values greater than 83%).

We then compare the results obtained on the Biometec and the public

Antimicrobial	**Maximum value of accuracy**	
	Whole dataset	**Respiratory data**
AMC	0,936	1,000
FEP	0,949	0,944
CAZ	0,979	1,000
IMI	0,941	1,000
AZT	0,981	1,000
CIP	0,933	0,929
CN	0,953	1,000
COL	0,987	NA
FOS	0,682	0,905
SXT	0,980	1,000

Table 6.1: Maximum values of accuracy obtained from the six ML models for the public dataset. Results are compared in the case when all 127 strains of the dataset or only the 37 strains of the respiratory tract have been considered. For the antimicrobial agent COL in the case of the respiratory tract strains, the accuracy cannot be evaluated since a single resistant strain is available in the dataset.

Antimicrobial	**Maximum value of accuracy**
IMI	0,920
AZT	0,970
CN	1,000
COL	0,963
FOS	0,833
SXT	0,917

Table 6.2: Maximum values of accuracy obtained from the six ML models for the Biometec dataset (57 strains).

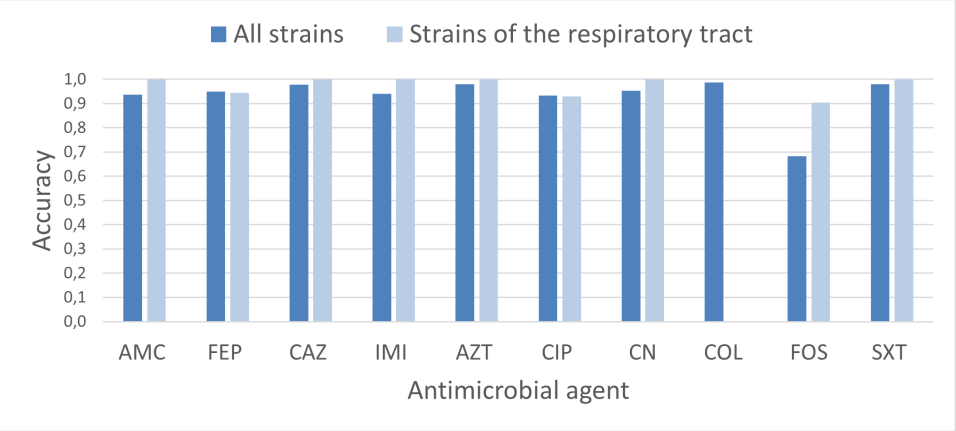


Figure 6.4: **Accuracy values for the public dataset.** Comparison of the maximum values of accuracy obtained when using all data of the public dataset or only those on bacterial strains taken from the respiratory tract.

dataset. In this case, the analysis is carried out on the six antimicrobial agents (IMI, AZT, CN, COL, FOS, and SXT) that are shared by the two datasets. The results are illustrated in Fig. 6.5, where, for each antimicrobial agent, we have reported the maximum value of accuracy obtained by applying the six ML methods to the two datasets. We note a higher value of accuracy for the FOS antimicrobial agent using the Biometec dataset, while for SXT antimicrobial agent a higher value of accuracy is obtained when the public dataset is used. For the other antimicrobial agents, similar values of accuracy are obtained in the two datasets. We also notice that, for all antimicrobial agents with the exception of FOS, the accuracy values are close to one.

The result obtained for the FOS antimicrobial agent is particularly interesting. The two datasets we have studied contain different genes. In particular, while the public dataset does not contain any *Fosfomycin*-type gene, the Biometec dataset includes one gene of this type. However, we have found that the presence/absence of this single gene is not enough to explain the different value of accuracy obtained for the two datasets. In fact, removing this gene from the Biometec datasets and applying the six ML methods to this reduced subset of data yields a very similar value of accuracy (in both cases the accuracy is between 0.703 and 0.833, depending on the method used). When Logistic regression, Gradient Boosting and Bag-

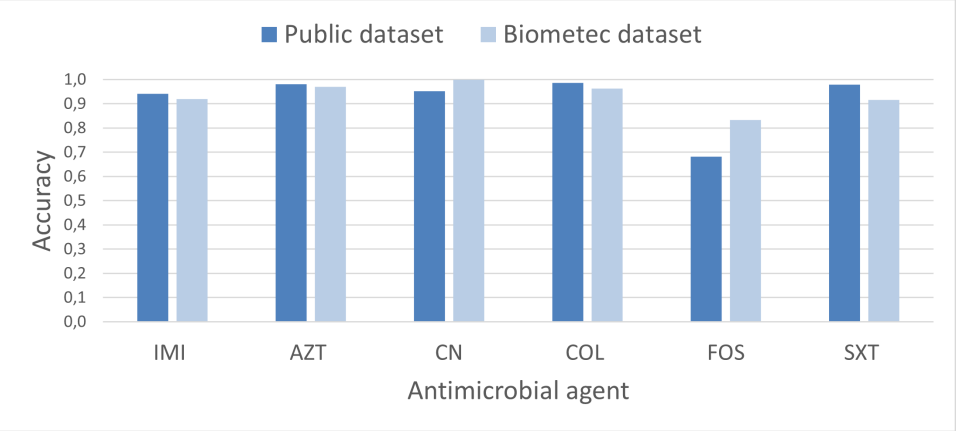


Figure 6.5: **Accuracy values for the public dataset and for the Biometec ones.** Comparison between the maximum value of accuracy obtained in the two datasets for six different antimicrobial agents (for the public dataset all 127 strains have been considered).

ging classifier are used, the same values of accuracy are obtained even if we consider the reduced dataset without the *Fosfomycin*-type gene, e.g. 0.708 for Logistic regression, 0.833 for Gradient Boosting, 0.750 for Bagging classifier. Instead, the accuracy values obtained with the other three methods differ slightly when we do not consider the *Fosfomycin*-type gene among the genes, in particular with Gaussian Naive Bayes we obtained 0.708 and 0.750, with k-nearest neighbors 0.833 and 0.708, with radius neighbors 0.750 and 0.792 respectively in the case of dataset with or without *Fosfomycin*-type gene.

This shows that the analysis performed is far from trivial. The low values of the Pearson correlation coefficient ρ observed for this antimicrobial agent (see Fig. 6.3) also suggest that other genes may be needed to predict resistance/susceptibility to this antimicrobial agent.

As discussed in Section 6.2, the analysis based on the Pearson correlation coefficient reveals a weak correlation between virulence genes and resistance to antimicrobial agents. For this reason, the analysis illustrated so far has been carried out only considering the data related to resistance genes. Here, we shortly comment on the results that are obtained by also including virulence genes in our analysis. The accuracy values obtained by considering data on both resistance and virulence genes are reported respectively in Ta-

Antimicrobial	**Maximum values of accuracy**	
	Resistance genes	**Resistance and virulence genes**
AMC	0,936	0,936
FEP	0,949	0,949
CAZ	0,979	0,979
IMI	0,941	0,971
AZT	0,981	0,981
CIP	0,933	0,956
CN	0,953	0,984
COL	0,987	0,987
FOS	0,682	0,758
SXT	0,980	0,960

Table 6.3: Maximum values of accuracy for the public dataset when using only resistance genes or when using both resistance and virulence genes.

ble 6.3 for the public dataset and in Table 6.4 for the Biometec dataset. We observe that, although for some antimicrobial agents (e.g. FOS, SXT, CZA) a higher value of accuracy is observed when data on the virulence genes are also included in the analysis, this is not true for all antimicrobial agents and there are several occurrences (e.g., AZT, COL, MEM, MEM/VAB, and AK) where a slightly lower value of accuracy is obtained. Based on these findings, a definitive answer to the question of whether or not the virulence genes should also be considered cannot be given. However, using data related only to resistance genes generally gives satisfactory performance in terms of values of accuracy that can be obtained.

Antimicrobial	Maximum values of accuracy	
	Resistance genes	**Resistance and virulence genes**
IMI	0,920	0,952
AZT	0,970	0,966
CN	1,000	NA
COL	0,963	0,957
FOS	0,833	0,900
SXT	0,917	1,000
CZA	0,846	1,000
MEM	0,889	0,870
MEM/VAB	0,920	0,905
AK	0,963	0,926

Table 6.4: Maximum values of accuracy for the Biometec dataset when using only resistance genes or when using both resistance and virulence genes. The results include the six antimicrobial agents common to the public dataset, and the four antimicrobial agents present only in the Biometec dataset.

Lastly, a separate analysis was made for the following antimicrobial agents: CZA, MEM, MEM/VAB and AK , which are present only in the Biometec dataset. These are, in fact, important antimicrobial agents for the clinical treatment of *Klebsiella pneumoniae*.

The results obtained by applying the six ML models of Section 4.2 to these antimicrobial agents are illustrated in Table 6.5. The values of accuracy are in a range from 0.654 to 0.963. The best performance is obtained with the use of logistic regression, k-nearest neighbors and bagging classifiers.

6.5 Final remarks

In this chapter, we addressed the prediction of AMR in *Klebsiella pneumoniae* through the application of various ML methods to genomic data. Six

Drug / ML Classifier	CZA	MEM	MEM/VAB	AK
Gaussian Naive Bayes	0,654	0,704	**0,920**	0,926
Logistic Regression	**0,846**	0,815	0,840	**0,963**
k-nearest Neighbors	**0,846**	**0,889**	0,840	**0,963**
Radius Neighbors	0,654	0,852	0,760	0,704
Gradient Boosting	0,731	**0,889**	0,840	**0,963**
Bagging	**0,846**	0,741	0,840	**0,963**

Table 6.5: Values of accuracy obtained by using different ML classifiers. The highest accuracy value(s) obtained for each antimicrobial agent are highlighted in **bold**.

algorithms were employed for this purpose: Gaussian Naive Bayes, Logistic Regression, k-Nearest Neighbors, Radius Neighbors, Gradient Boosting, and Bagging Classifier. The performance of these models was evaluated using two distinct datasets, the Biometec dataset and a public dataset, which differ in the number of *K. pneumoniae* strains, their biological sources (blood, urine, respiratory tract), and the set of genes included as features. By comparing their performances, we aimed to assess the predictive capacity of different genomic features and evaluate the robustness of ML approaches in this context.

All six models demonstrated satisfactory predictive performance, even when only resistance genes were considered, excluding virulence genes that did not contribute meaningfully to classification accuracy. For the Biometec dataset, the highest accuracy (above 90%) was achieved using the KNN algorithm. Comparable performance was observed for most antibiotics common to both datasets, with the exception of Fosfomycin (FOS), where discrepancies were attributable to the presence or absence of the corresponding resistance gene in each dataset. Notably, no direct one-to-one correspondence between individual genes and specific antimicrobial resistances emerged, suggesting that resistance patterns arise from complex, non-linear interactions among multiple genomic factors, an area where ML methods prove particu-

larly effective.

The ML models performed well in both datasets, despite substantial differences in their structure. The public dataset, though containing a larger number of strains (127) and a smaller number of resistance-associated genes (9), yielded comparable results to the Biometec dataset (57 strains, 34 genes). This indicates that the models were capable of generalizing even with a limited number of features or samples. Similarly, when the analysis was restricted to specific subsets, such as strains from the respiratory tract, the models maintained good predictive performance. These findings highlight the robustness of the proposed ML approaches and their ability to capture key genomic determinants of resistance across different data configurations.

Nevertheless, several limitations must be acknowledged. First, the relatively small size of the available datasets may limit the models' generalizability to unseen data, potentially increasing the risk of overfitting. Data imbalance, observed in both datasets, represents another challenge: when one class (e.g., resistant strains) predominates, models may become biased, reducing their predictive accuracy for the minority class. Future studies could employ advanced resampling techniques or expand the datasets to mitigate these issues. Moreover, model performance varied across antibiotics, indicating that the optimal ML algorithm may depend on the specific resistance mechanism under study.

Although computational times were negligible, practical constraints persist. In particular, the acquisition of genomic data through sequencing and subsequent bioinformatic processing remains time-consuming, limiting the immediate clinical applicability of these predictive models. Nevertheless, the present work provides a preliminary yet promising indication of the potential of data-driven methods for AMR prediction.

From a broader perspective, the use of ML in AMR research offers several advantages. It enables the efficient analysis of large genomic datasets and the discovery of complex patterns that would be difficult to detect using conventional statistical approaches. Such models can support the early identification of resistant strains, facilitating more informed antibiotic prescription and contributing to the reduction of inappropriate antimicrobial use. Furthermore, ML methods can aid in the discovery of novel antibi-

otics by screening extensive chemical or genomic databases, as well as in the personalization of antimicrobial therapies tailored to specific bacterial or patient profiles.

As a future direction, the integration of phenotypic data (such as MIc values) could improve model interpretability and accuracy. Since the relationship between genotypic variations and phenotypic resistance is often non-linear, combining both types of information would allow for a more comprehensive understanding of resistance mechanisms. Additionally, incorporating contextual factors such as geographical origin, clinical setting (e.g., hospital- vs. community-acquired infections), prior antibiotic exposure, or other demographic information could further enhance predictive performance.

In summary, these results demonstrate that ML approaches can effectively predict antimicrobial resistance in *K. pneumoniae* using genomic data alone, even in the presence of limited and heterogeneous datasets. While further validation on larger and more diverse collections is required, these results underscore the methodological and conceptual potential of ML as a tool for advancing AMR surveillance and research.

Chapter 7

Missing Data Reconstruction

This chapter describes the methodological framework and experimental results related to the reconstruction of missing data. The goal of this analysis is to assess the performance of different machine learning approaches in recovering both binary and continuous values under various missingness scenarios. We begin by briefly outlining the theoretical background of the missing data problem, introducing the main types of missingness mechanisms and their implications for data analysis and model reliability. Subsequently, we detail the preprocessing procedures applied to the datasets, including the artificial generation of missing values according to both random (*MCAR*) and structured non-random (*MNAR*) mechanisms. The following sections describe the configuration of the machine learning models employed for imputation and the evaluation metrics used to quantify reconstruction accuracy. Results are then presented separately for binary and continuous datasets, allowing a comparative interpretation of model performance across data types.

7.1 Missing data problem

The term *missing data* or *missing values* refers to the absence in a dataset of one or more values relating to an observed variable, e.g. in the case when data have not been stored [51, 87]. This can happen for various reasons, such as measurement errors, failures in acquisition systems, or human transcription errors. This problem is often present in medical datasets, so it is necessary to manage it as it could significantly complicate the statistical

analysis and alter the results.

Missing data can be classified into three categories [96, 33, 73]:

- *Missing Completely at Random (MCAR)*: describes a situation in which the likelihood of missing data is entirely unrelated to the values of any variables in the dataset, whether observed or unobserved. In other words, every observation has the same probability of missing data, regardless of its characteristics. This means that the missingness occurs purely by chance, and not due to any underlying pattern or relationship in the data. Because the mechanism causing the missing data is unrelated to the data itself, analyzes performed on the subset of complete cases will still yield unbiased and valid results. However, the main drawback is the reduction in sample size, which can affect the statistical analysis.

- *Missing at Random (MAR)*: refers to a missing mechanism in which the probability of a data point is related to the values of the observed data, but not to the unobserved (missing) data itself. In other words, the missingness is systematic but explainable using the information already available in the dataset. Under MAR, the missing data are not random in the strictest sense: they are conditional on observed variables. As a result, simply ignoring missing data (e.g., via complete case analysis) can introduce bias. However, with appropriate statistical techniques, such as multiple imputation or maximum likelihood estimation, the missingness can be properly accounted for, and valid inferences can still be made.

- *Missing Not a Random (MNAR)*: refers to situations where the probability that data are missing is systematically related to both the observed and the unobserved values. In this case, the missingness depends directly on the data that are themselves missing, meaning that the absence of data is not random but instead driven by specific, often latent, characteristics of the data. This dependency introduces significant challenges for analysis, as standard imputation or modeling techniques that assume randomness in the missingness mechanism (e.g., MCAR or MAR) can lead to biased or misleading results. The inherent difficulty with MNAR data lies in the fact that the causes of

missingness cannot be inferred from the observed data alone, making it challenging to model or correct for the missingness mechanism. As a result, MNAR often requires specialized methods, strong assumptions, or external data sources to handle appropriately.

It is also important to recognize and categorize patterns of missing data within a dataset, as this can provide insights into the underlying mechanisms and help guide the choice of appropriate imputation or analysis methods [63, 95, 12]. Some commonly identified missing data patterns include (Fig. 7.1):

- *univariate pattern*: missing values occur only in a single variable, with all other variables fully observed. There is no apparent relationship between the missingness and the other variables in the dataset;
- *multivariate pattern*: missing values are present across multiple variables, but the pattern of missingness does not follow a specific sequence. The missingness may still be unrelated to the values of other variables;
- *monotone pattern*: the missing data follow a structured sequence, often observed in longitudinal or ordered data. For example, if a variable is missing for a given observation, all subsequent variables (in order) are also missing for that same observation. This often arises in studies with dropout or time-series structures;
- *general (or arbitrary) pattern*: missing values occur without any apparent order or structure. The missingness is scattered across the dataset, and it may be random or not. This is the most complex scenario to handle, especially if the mechanism is not MCAR;
- *file-matching pattern*: this occurs when the dataset results from merging two or more sources (e.g., datasets from different studies or institutions), where some variables are present in one dataset but missing in another. As a result, missingness aligns with the origin of the data rather than with individual measurements.

Understanding the structure of missing data is crucial for choosing suitable imputation techniques and ensuring that the analysis does not produce

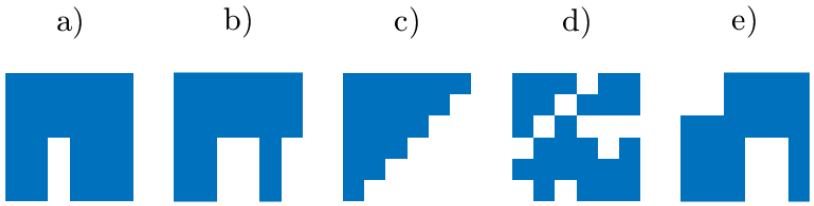


Figure 7.1: **Different pattern of missing values:** (a) univariate: only one variable has missing values; all other variables are fully observed; (b) multivariate: missing values in a number of variables; the pattern of missingness does not follow a particular order; (c) monotone: missing data with a systematic pattern, frequently in ordered or longitudinal data; (d) general: missing values occur without any apparent order or structure; (e) file-matching: occurs when two or more sources are combined to create a dataset; some variables are present in one dataset but not in another.

biased or invalid results.

Different approaches can be adopted to analyze data containing missing data, such as eliminating the entire sample containing missing data from the analysis (*listwise deletion*). However, such an approach is only possible in the presence of MCAR-type missing data and if the dataset remains consistent and significant even after eliminating such incomplete samples. In the case in which this approach is not applicable, a possible solution is to replace the missing data with estimated values, i.e., data imputation.

While MCAR reconstruction is an essential benchmark for assessing the robustness of imputation algorithms, several AMR-specific studies have shown that real-world AMR data rarely conform to this assumption. In large surveillance or genomic datasets, data missingness often follows structured, process-driven, or selection-dependent patterns. Some studies demonstrated that public repositories, such as PATRIC, exhibit systematic missingness resulting from unbalanced species representation, geographic bias, and incomplete phenotypic testing [91], and reported that multi-label genomic datasets contain missing labels directly linked to laboratory test design, as susceptibility is measured only for a subset of antibiotics [107]. A recent study [108] further emphasized that the Pfizer ATLAS dataset includes MNAR components, which combine multivariate dependencies, file matching incon-

sistencies, and a fraction of random omissions. Therefore, while the evaluation on MCAR data is crucial to ensure methodological reliability under controlled conditions, a comprehensive analysis must also take into account MNAR patterns, which more accurately reflect the mechanisms underlying the generation of real-world AMR data. In this study, we therefore extend the evaluation to MNAR scenarios, integrating experiments conducted on MCAR data. In the following sections, some imputation techniques will be explained in detail.

7.2 Preprocessing of the datasets

To evaluate the robustness and effectiveness of imputation strategies, it is necessary to test them under controlled missing data conditions. Since the datasets we used are complete or already preprocessed, we artificially introduced missing values according to well-defined mechanisms. This allows a systematic assessment of model performance under different missing data scenarios, ensuring reproducibility and comparability of the results. In particular, two complementary approaches were implemented: the generation of MCAR data, where the probability of a value being missing is independent of any observed or unobserved variable, and the generation MNAR data, where the missingness mechanism depends on the data structure itself, introducing non-random and potentially correlated patterns of absence. The following subsections describe the methodologies and the corresponding Python implementations adopted for each case.

7.2.1 Random value removal

In order to create missing data artificially, we chose to remove data, following a completely random missing data mechanism, starting from complete datasets. For this purpose, a function was implemented in Python, using the pandas and NumPy libraries. This function has two input parameters: the dataset from which we will remove some elements and the percentage of elements to remove. A Boolean mask is then made by generating a Boolean vector of false elements with the same size as the dataset, in which a number equal to the number of elements to remove is set to true, in random positions. This vector is then reformatted to fit the shape of the dataset in

order to set the values corresponding to the true values of the mask to NaN. The function returns the new dataset with missing data and the applied mask, in order to keep track of the changes and facilitate comparison with the original dataset.

7.2.2 Not a random value removal

For MNAR data generation, we implemented a Python function to introduce structured, non-random missingness patterns. The procedure leverages the pandas and NumPy libraries, enabling precise control over the missingness structure while maintaining a predefined global missing rate. Three main types of patterns can be simulated: *column-wise patterns*, where missing values are generated along selected features according to a two-state Markov process defined by transition probabilities; *row-wise patterns*, which apply the same logic horizontally to simulate dependent missing segments within observations; and *block-shaped patterns*, in which rectangular regions of the dataset are removed to emulate file-matching or multi-source data scenarios. A small proportion of random noise can also be introduced to reproduce mixed MNAR-MCAR conditions. A fine-tuning step ensures that the overall proportion of missing values precisely matches the target missing rate, while a safeguard prevents complete feature removal by enforcing a minimum percentage of observed entries per column. The function outputs both the dataset with missing values and the corresponding Boolean mask, facilitating reproducible evaluation of imputation performance.

7.3 Machine Learning models setting

As reported in Section 4.3, we focus on four techniques: KNN, RF, MICE and KMI. All of these algorithms were implemented using Python.

Starting from the first, KNN imputation has been used for both types of datasets, binary and continuous. To implement the KNN imputation algorithm, the `KNNImputer` class from the `sklearn.impute` module of Python's scikit-learn library was used. `KNNImputer` uses a multivariate imputation algorithm, therefore it exploits the entire set of available features to estimate missing values, unlike univariate imputation, in which one feature is considered at a time. Inside the Python function, the parameters `n_neighbors`,

`weights` and `metric` are specified. The `n_neighbors` parameter is used to determine the number of samples close to the sample to be imputed that must be taken into account for the calculation of missing values, `weights` establishes the weight that each sample must have in the calculation, and `metric` chooses the distance metric to apply to identify the samples closest to the one under consideration.

The major advantage of the KNN method is that it is simple, effective for small datasets and preserves the local structure of the dataset. However, it is computationally expensive for large datasets and is sensitive to the choice of k.

The second method, RN, is used inside an iterative inputer, implemented in Python using in this case the `IterativeImputer` class from the `sklearn.impute` module of Python's scikit-learn library. In the Python function we need to specify the parameter `estimator`, which refers to the regressor model to use to predict the missing data, in this case `RandomForest Regressor`. In turn, the random forest regressor has specific parameters that can be defined, such as the number of trees (`n_estimators`) or a seed (`random_state`) to ensure reproducibility of the results.

The main advantages of the RF method are that it handles both numerical and categorical data, captures complex interactions, and provides robust imputations, while the main limitations it can be computationally demanding and may overfit to noisy data.

In this work, the RF algorithm has been used only for the imputation of binary data.

MICE model was implemented in Python using an `IterativeImputer` with `estimator= BayesianRidge` for the binary dataset and `estimator= RandomForestRegressor` for the continuous dataset. Although this algorithm introduces significant improvements over traditional models, it can be computationally demanding, especially with very large data sets. Furthermore, as mentioned, it operates under the assumption that the missing data are of the MAR or MCAR type.

Finally, in the case of datasets containing continuous values, the KMI algorithm can be used to estimate missing data.

In Python, the algorithm was implemented using the `KMeans` class from the `sklearn.cluster` module, on a copy of the dataset where missing val-

ues were replaced with the mean of each feature. The elbow method was implemented using the `KneeLocator` function from the `kneed` library.

7.4 Results

In this section, we present the results of our tests on the reconstruction of missing data across some of the biological datasets described in Chapter 3. In particular, for this work we used the public (3.1.2) and the synthetic (3.1.3) binary datasets, and the five continuous sub-datasets described in Section 3.2. Each dataset was analyzed using appropriate evaluation metrics based on the nature of the data, binary or continuous.

For binary datasets, which typically represent the presence or absence of antimicrobial resistance traits or gene features, we assess reconstruction performance using five standard classification metrics: accuracy, precision, recall, F1-score and the AUC (5.1). In the case of continuous datasets, such as those containing quantitative measurements or expression levels, we evaluate the quality of the imputation using NRMSE and NMAE (5.2).

Together, these evaluation criteria provide a comprehensive view of how well the proposed methods reconstruct missing values under varying data types and conditions. The following subsections detail the experimental outcomes, highlighting comparative performance across different imputation approaches.

7.4.1 Binary Datasets

The first step in our workflow involved pre-processing the publicly available dataset to extract only the information relevant to our study, specifically the presence or absence of antimicrobial resistance and virulence genes in different bacterial strains. This data cleaning and selection process allowed us to isolate a focused subset of features suitable for evaluating missing data reconstruction methods. From this curated dataset, a second dataset was derived as explained in 3.1.3, which includes a synthetic variant used to assess performance under controlled conditions.

Following the definition of the datasets, we established the procedure for dividing them into training and test sets, which is necessary to evaluating the performance of the imputation methods. To ensure robust and general-

izable results, we employed a cross-validation strategy that involves multiple repetitions on different train-test splits. This approach allows imputation models to be trained and evaluated in various subsets of the data, mitigating the risk of overfitting to a particular sample and enabling a more reliable assessment of the performance of each method.

To ensure robust evaluation, we implemented cross-validation using the `KFold` class from the `sklearn.model_selection` module in Python. This method partitions the dataset into K equally sized folds, and the imputation process is repeated K times: at each iteration, one fold is used as the test set while the remaining $K-1$ folds serve as the training set. The choice of K was guided by dataset size: for the smaller public dataset we set $K = 5$, while for the larger synthetic dataset we used $K = 11$, ensuring sufficient variability across splits.

This methodological framework allows for a statistically evaluation of the imputation techniques across varying data configurations, improving the reliability and reproducibility of our findings.

In addition to the K parameter, it is also necessary to set the parameters related to the individual imputation techniques for both datasets. In particular, the parameters that guarantee the best imputation performance, reducing the computational cost as much as possible, were chosen. For the KNN method, the value of `n_neighbors` was set to 5 for both datasets, being the value for which the mean square error (MSE) reaches its minimum (≈ 0.01), while for `weights` and `metrics` we assigned a uniform weight to all neighbors and used the Euclidean distance to determine them. For the iterative imputation method with RF regressor, `n_estimators=5` (since it guaranteed the greatest reduction in MSE) and `max_iter=6` (since increasing further results in an average larger MSE) were set. Finally, for the MICE imputation, the key parameter is the number of imputations to be performed and, consequently, the number of datasets generated. In our case, it was set to `max_iter=6` for both datasets. These values were chosen to balance computational efficiency and accuracy, as further increasing the number of imputations yielded minimal improvement in performance.

Having discussed the experimental settings, we now proceed to analyze the results obtained from the first dataset: the public binary dataset. This dataset comprises 127 samples (representing bacterial strains) and 25 bi-

nary features, corresponding to the presence or absence of virulence and resistance genes. Given its relatively small size, the dataset is less prone to overfitting and less affected by noise, which generally makes the imputation task more manageable compared to larger, more complex datasets. In total, the dataset contains 3.175 values. Among these, 1.413 values (44.5%) indicate the presence of a gene (encoded as 1), while 1.762 values (55.5%) represent the absence of a gene (encoded as 0). These proportions suggest that the dataset is reasonably balanced, with a slight prevalence of absent genes.

The performance of the imputation algorithms is evaluated as the percentage of missing data increases. Table 7.1 summarizes the performance of three imputation methods discussed in Section 7.3, KNN, RF, and MICE, evaluated using accuracy and AUC on increasing levels of missing data (from 10% to 50%). These metrics allow us to assess how well the imputed data reflect the original structure across different levels of the data incompleteness.

The results reveal that MICE method consistently outperforms the other two methods in terms of accuracy and AUC, across all missing data rate values, indicating its robustness in handling data sparsity. RF shows competitive performance at lower missing rates (below 20%), but tends to degrade more rapidly as missingness increases. KNN, while simple and computationally efficient, yields lower overall scores, especially in AUC, and is more sensitive to the percentage of missing values. The comparison highlights the importance of choosing advanced and model-aware imputation techniques such as MICE, especially when working with datasets affected by moderate to severe missingness.

For all experiments on both public and synthetic datasets, each imputation scenario was repeated over 10 independent simulation runs. The values reported in all tables therefore represent the average performance across these 10 iterations, providing a more stable and statistically reliable estimate of the behaviour of each method.

To evaluate the effectiveness and scalability of the imputation methods on larger datasets, we constructed a synthetic dataset derived from the structure of the original public dataset. The goal was to preserve the statistical properties of the original data, and in particular, the correlation structure

among features, which plays a crucial role in realistic data reconstruction tasks. By preserving this correlation structure, the synthetic dataset replicates the dependencies among features found in the real data, allowing us to more accurately assess the generalizability of imputation strategies. The resulting dataset contains 25.000 values, of which 12.162 (49%) indicate the presence of genes (value 1) and 12.838 (51%) indicate the absence (value 0). Thus, the dataset remains well balanced, similar to the original one.

The performance of the imputation methods applied to this larger, structurally realistic dataset is summarized in Table 7.2, where accuracy and AUC values are reported across varying levels of missing data. The table highlights how each model handles the progressive degradation of data and helps us evaluate their robustness and reliability. The MICE method consistently achieves the highest accuracy and AUC values across all missing data percentages. Its performance remains remarkably stable, particularly up to 30% missing data, with only a moderate decline beyond that. This demonstrates MICE's strong ability to reconstruct missing values even under challenging conditions. RF performs moderately well, ranking second in most cases. Its results are relatively close to those of MICE up to 30% missing data but show a slightly steeper decline at higher missing rates, especially in AUC. KNN exhibits the lowest performance among the three methods. Its accuracy and AUC steadily drop as the missing data rate increases, suggesting that it is more sensitive to sparsity and less robust when large portions of data are unavailable.

Accuracy and AUC generally follow the same trend for all methods, indicating that both metrics are reliable indicators of imputation performance in this context. AUC values are consistently slightly higher than the accuracy for all the methods, highlighting their ability to preserve class ranking even when some fine-grained predictions might be incorrect.

We can remark that with missing data at 10%, MICE already leads with 0.8340 accuracy and 0.8287 AUC, setting a strong baseline. Even with missing data at 50%, MICE maintains 0.7692 accuracy and 0.7682 AUC, outperforming the other two methods by a noticeable margin. The performance gap between the MICE method and the other methods becomes more pronounced as the missing percentage increases, underscoring its robustness in more challenging scenarios.

Beyond accuracy and AUC, the remaining evaluation metrics, precision, recall, and F_1-score, exhibit analogous behaviour. In both public and synthetic datasets, MICE consistently achieves the best or near-best values across all levels of missingness, while RF typically ranks second and KNN remains the least effective method. These results, reported in Tables 7.1 and 7.2, confirm that the superiority of MICE is not limited to a specific performance indicator but is reflected across all classification metrics.

When analyzing the MNAR scenario, reported in Tables 7.3 and 7.4, the same overall trends are observed. MICE remains the most robust method across all missing levels, maintaining high accuracy, precision, recall, F_1-score, and AUROC even as MNAR-induced missingness increases. RF delivers intermediate performance, while KNN is the most affected by MNAR mechanisms, especially in high-dimensional datasets. These trends are also clearly visible in Figs. 7.2-7.3 , which illustrate the trajectories of Accuracy, F_1-score and AUC under MNAR missingness for both the public and synthetic datasets.

7.4.2 Continuous datasets

Following the evaluation of the performance of imputation on binary datasets, we extended our analysis to continuous datasets that capture the relative abundance of ARGs across five distinct environmental habitats: soil, water, sediment, particulate matter, and dust. These datasets differ significantly in terms of sample size, feature dimensionality, and data distribution characteristics, offering a more comprehensive and challenging testbed for imputation algorithms. Each sub-dataset reflects real-world variability in environmental sampling and sequencing, which is common in AMR surveillance studies.

As discussed in the previous section, for the evaluation of imputation performance on continuous datasets, we employed two commonly used error-based metrics: NMAE and NRMSE. Table 7.6 presents the imputation results for varying levels of missing data, ranging from 10% to 50%, across three imputation algorithms: KNN, KMI, and MICE. This comparative analysis allows us to assess each method's robustness and accuracy as the percentage of missing data increases, offering valuable insights into their applicability to real-world environmental AMR datasets.

As with binary datasets, the application of imputation techniques to con-

Missing (%)	ML Model	Accuracy	Precision	Recall	F_1-score	AUC
10	KNN	0.886 278	0.871 068	0.813 480	0.861 156	0.881 204
	RF	0.900 028	0.890 119	0.888 801	0.894 258	0.901 978
	MICE	**0.926 203**	**0.926 237**	**0.918 477**	**0.918 222**	**0.925 182**
20	KNN	0.881 785	0.874 201	0.813 497	0.858 580	0.877 507
	RF	0.907 508	0.893 881	0.885 577	0.893 916	0.907 625
	MICE	**0.923 108**	**0.937 696**	**0.886 188**	**0.910 771**	**0.919 401**
30	KNN	0.873 989	0.881 614	0.797 163	0.850 773	0.868 675
	RF	0.897 872	0.884 186	0.872 883	0.883 502	0.895 325
	MICE	**0.907 646**	**0.906 188**	**0.886 293**	**0.896 353**	**0.905 702**
40	KNN	0.865 015	0.876 000	0.787 098	0.843 317	0.858 915
	RF	0.889 708	0.879 264	**0.862 744**	0.874 886	0.887 636
	MICE	**0.898 815**	**0.909 192**	0.861 212	**0.884 581**	**0.894 725**
50	KNN	0.849 821	0.858 519	0.766 754	0.823 785	0.843 616
	RF	0.881 256	0.878 343	**0.841 605**	**0.864 370**	**0.878 317**
	MICE	**0.884 551**	**0.904 569**	0.824 931	0.860 212	0.877 560

Table 7.1: Imputation performances metrics for different values of the percentage of missing values (MCAR) in the binary public dataset. For each missing percentage, the best-performing method is highlighted in **bold**.

tinuous data also requires careful tuning of algorithm-specific parameters for each dataset to optimize performance. In particular, for the KNN imputation method, three key parameters were configured: `n_neighbors`, which specifies the number of nearest samples to consider when imputing a missing value; `weights`, which determines how much influence each neighbor has on the imputed value and can be set to 'uniform' (equal weight for all neighbors) or 'distance' (neighbors closer in feature space have more influence); `metric`, that defines the distance function used to identify the nearest neighbors (e.g. Euclidean distance). The values of `n_neighbors` were selected individually for each dataset and are reported in Table 7.5. Regarding the weights parameter, we adopted the criterion of assigning greater influence

Missing (%)	Method	Accuracy	Precision	Recall	F_1-score	AUC
10	KNN	0.791 649	0.802 452	0.737 394	0.773 427	0.790 291
	RF	0.802 200	0.798 531	**0.789 681**	0.796 813	0.803 427
	MICE	**0.810 396**	**0.823 643**	0.765 561	**0.797 184**	**0.808 515**
20	KNN	0.779 304	0.791 426	0.721 278	0.760 275	0.777 373
	RF	0.795 000	0.788 399	**0.773 721**	0.784 869	0.793 937
	MIC	**0.800 449**	**0.801 992**	0.767 461	**0.788 866**	**0.799 814**
30	KNN	0.759 327	0.782 620	0.686 676	0.734 406	0.757 348
	RF	0.778 533	0.763 807	**0.763 530**	0.768 021	0.777 976
	MICE	**0.792 815**	**0.807 801**	0.741 978	**0.779 979**	**0.792 031**
40	KNN	0.737 667	0.750 224	0.662 681	0.712 044	0.735 810
	RF	0.761 350	0.748 920	**0.750 744**	0.752 058	0.761 097
	MICE	**0.778 545**	**0.787 393**	0.732 351	**0.761 577**	**0.777 141**
50	KNN	0.705 937	0.707 828	0.625 200	0.670 592	0.702 787
	RF	0.740 880	0.725 134	**0.734 457**	0.730 157	0.740 549
	MICE	**0.759 204**	**0.767 901**	0.707 419	**0.735 824**	**0.756 389**

Table 7.2: Imputation performances metrics for different values of the percentage of missing values (MCAR) in the binary synthetic dataset. For each missing percentage, the best-performing method is highlighted in **bold**.

to closer neighbors in the imputation process, under the assumption that nearby instances in feature space carry more relevant information for estimating missing values. This approach corresponds to setting the parameter `weights=‘‘distance"` in the KNN imputer. Our experiments revealed that this configuration led to marginally better performance (with improvements ranging from approximately 1% to 2.5% in terms of NRMSE) for the particulate matter (PM) and sediment (SD) datasets. For the remaining three datasets (soil, water, and dust), however, the choice between ‘uniform’ and ‘distance’ weighting showed no substantial impact on imputation accuracy. These findings underscore the importance of dataset-specific parameter tuning to maximize the effectiveness of imputation methods in heterogeneous

Missing (%)	Method	Accuracy	Precision	Recall	F_1-score	AUROC
10	KNN	0.902 630	0.881 357	0.865 527	0.878 170	0.897 934
	RF	0.932 857	0.900 248	0.907 462	0.917 287	0.932 089
	MICE	**0.949 132**	**0.916 780**	**0.956 535**	**0.936 475**	**0.950 467**
20	KNN	0.831 538	0.808 137	0.798 214	0.802 623	0.827 486
	RF	0.896 523	0.862 133	0.900 879	0.891 196	0.897 697
	MICE	**0.928 400**	**0.911 300**	**0.932 164**	**0.925 521**	**0.929 009**
30	KNN	0.838 462	0.843 747	0.758 052	0.806 622	0.830 970
	RF	0.880 649	0.874 991	0.840 912	0.866 311	0.877 475
	MICE	**0.887 962**	**0.891 094**	**0.860 835**	**0.874 155**	**0.884 991**
40	KNN	0.824 231	0.843 505	0.741 302	0.800 733	0.820 288
	RF	0.841 000	0.848 614	0.810 804	0.825 913	0.839 443
	MICE	**0.860 754**	**0.869 762**	**0.824 599**	**0.843 757**	**0.857 972**
50	KNN	0.815 692	0.834 476	0.702 269	0.778 510	0.808 153
	RF	0.829 103	0.825 698	0.793 490	0.807 429	0.826 701
	MICE	**0.866 987**	**0.860 100**	**0.809 657**	**0.847 346**	**0.864 086**

Table 7.3: Imputation performances metrics for different values of the percentage of missing values (MNAR) in the binary public dataset. For each missing percentage, the best-performing method is highlighted in **bold**.

biological and environmental data contexts.

Dataset	SL	WT	SD	PM	DT
k	8	10	2	5	4

Table 7.5: Number k of adjacent samples used for the imputation in the 5 datasets with continuous values.

For the KNN method, the `metric` parameter was set to `nan_euclidean`, which computes pairwise Euclidean distances while appropriately handling missing values. This metric ensures that only the features observed in both

Missing (%)	Method	Accuracy	Precision	Recall	F_1-score	AUROC
10	KNN	0.740 784	0.691 907	0.547 619	0.611 845	0.701 824
	RF	0.755 860	0.694 437	0.671 848	0.693 140	0.746 917
	MICE	**0.792 018**	**0.778 294**	**0.699 953**	**0.739 365**	**0.780 291**
20	KNN	0.717 856	0.680 604	0.560 826	0.617 424	0.694 826
	RF	0.723 758	0.678 521	**0.658 990**	0.677 446	0.717 113
	MICE	**0.762 857**	**0.760 307**	0.669 454	**0.715 400**	**0.754 588**
30	KNN	0.722 454	0.693 891	0.582 271	0.645 256	0.708 390
	RF	0.717 891	0.656 388	**0.668 302**	0.675 492	0.716 203
	MICE	**0.760 990**	**0.757 931**	0.657 869	**0.711 797**	**0.749 684**
40	KNN	0.716 802	0.711 184	0.593 927	0.652 734	0.707 179
	RF	0.726 593	0.683 165	**0.692 972**	0.703 135	0.726 087
	MICE	**0.772 308**	**0.772 853**	0.689 257	**0.736 052**	**0.768 059**
50	KNN	0.688 224	0.671 752	0.567 395	0.628 464	0.681 732
	RF	0.700 012	0.655 826	**0.670 377**	0.676 568	0.699 495
	MICE	**0.743 378**	**0.741 829**	0.655 150	**0.698 572**	**0.736 854**

Table 7.4: Imputation performances metrics for different values of the percentage of missing values (MNAR)in the binary synthetic dataset. For each missing percentage, the best-performing method is highlighted in **bold**.

samples contribute to the distance calculation, allowing for meaningful similarity comparisons even when some data points are incomplete.

In the case of the MICE method, two key parameters must be configured, one related to the underlying regression model (in our case, random forest) and one specific to the imputation process itself. For the regression model, the most relevant parameter is `n_estimators`, which determines the number of decision trees used in the RF. Through empirical testing, we found that `n_estimators=10` provides a good balance between the accuracy of imputation and computational efficiency. Increasing the number of trees beyond this threshold did not significantly improve performance while substantially increasing computation time.

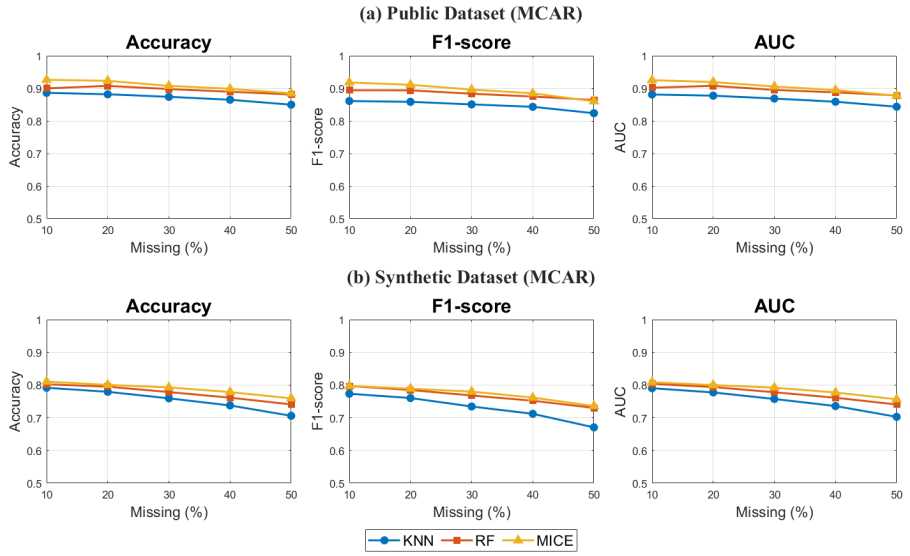


Figure 7.2: **Imputation performance under MCAR missingness for binary classification tasks:** the plots report Accuracy, F1-score, and AUC as a function of the percentage of missing data (10–50%) for three imputation methods (KNN, RF, and MICE). Results are shown for (a) the public dataset and (b) the synthetic dataset. Higher values indicate better classification performance after imputation.

For the MICE algorithm itself, the parameter `max_iter` controls the number of iterative cycles used to refine the imputed values. Consistent with our approach in the binary data experiments, we set `max_iter=5` across all continuous datasets. Additional iterations did not yield meaningful performance gains and only prolonged the runtime, confirming this setting as an efficient compromise.

As for the KMI, the main parameter of interest is the number of clusters. This value was dynamically determined using the elbow method, as described earlier. This automated selection ensures that the number of clusters reflects the inherent structure of each dataset, helping to optimize imputation quality without manual tuning.

The analysis of the continuous ARG datasets (SL, WT, SD, PM, and DT), whose sizes range from 28 to 691 samples and from 28 to 102 features (Table 3.5), shows that all three imputation approaches, KNN, KMI, and MICE, achieve overall low reconstruction errors across all habitats. As re-

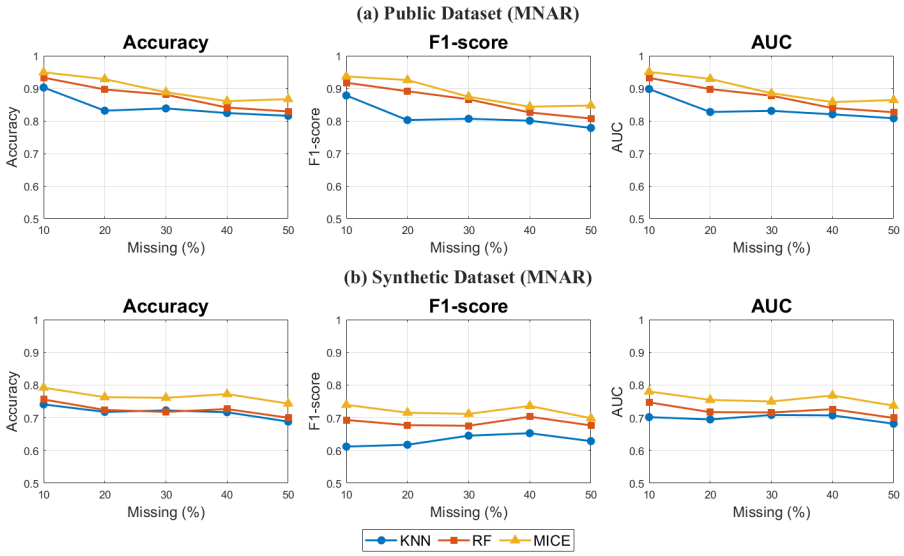


Figure 7.3: **Imputation performance under MNAR missingness for binary classification tasks:** the plots report Accuracy, F1-score, and AUC as a function of the percentage of missing data (10–50%) for three imputation methods (KNN, RF, and MICE). Results are shown for (a) the public dataset and (b) the synthetic dataset. Higher values indicate better classification performance after imputation.

ported in Tables 7.6–7.7, both NMAE and NRMSE values remain in a very small numerical range for the SL, WT, and SD datasets, and only moderately higher for the smaller PM and DT datasets. This indicates that the underlying continuous structure of gene-abundance values is well preserved even when a substantial fraction of entries is missing. Furthermore, the three methods exhibit remarkably similar trends across missing-value levels, with only modest differences between them. In particular, MICE often attains the lowest error for several dataset–metric pairs, especially under MCAR conditions, whereas KNN frequently provides competitive or slightly better values on specific datasets. KMI typically performs close to the other two methods but only rarely achieves the best score, suggesting a slightly inferior robustness in this setting. Nevertheless, the magnitude of these differences remains small, confirming that all methods provide a reliable reconstruction of continuous ARG profiles.

A second important observation is the stability of the reconstruction er-

ror with respect to the percentage of missing data. Error values vary only minimally from 10% to 50% missingness, with no substantial degradation in any dataset or metric. This consistent behavior across methods and habitats indicates that the difficulty of the imputation task does not increase significantly even when half of the dataset is removed. Because of this strong stability, it becomes meaningful to focus on a single representative missingness level in order to analyze in greater detail how performance differs across datasets of different sizes. For this reason, the subsequent bar-plot analysis (Fig. 7.4 and 7.5) examines the 30% missingness scenario, highlighting how the average reconstruction error changes with the number of samples and features in each habitat. These figures confirm that larger datasets (e.g., SL and WT) obtain the lowest imputation errors, whereas very small datasets (PM, DT) naturally exhibit higher variability and larger min–max ranges across methods, although still within a generally acceptable error margin.

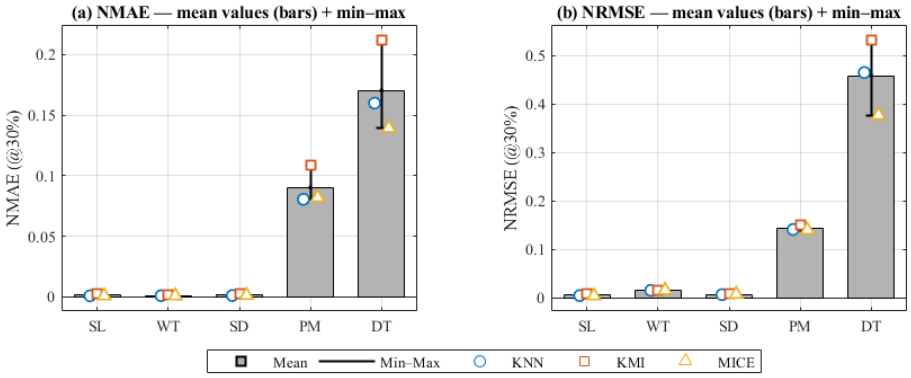


Figure 7.4: **Aggregated imputation performance at 30% missingness across all continuous datasets, in case of MCAR missing data:** for each dataset, the bar indicates the mean error over the three imputation methods, the whiskers show the min–max variability, and the overlaid markers represent the individual performances of KNN, KMI, and MICE. The figure provides a compact comparative view of the variability and absolute performance of the methods.

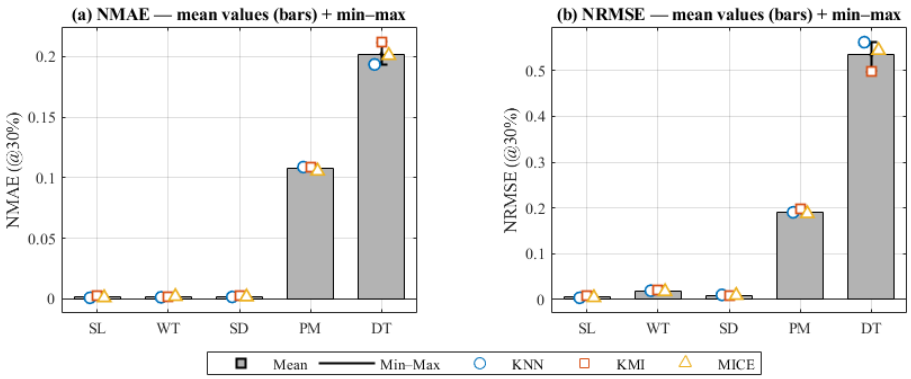


Figure 7.5: **Aggregated imputation performance at 30% missingness across all continuous datasets, in case of MNAR missing data:** for each dataset, the bar indicates the mean error over the three imputation methods, the whiskers show the min–max variability, and the overlaid markers represent the individual performances of KNN, KMI, and MICE. The figure provides a compact comparative view of the variability and absolute performance of the methods.

7.5 Final remarks

In this chapter, we systematically assessed the impact of different imputation strategies on the reconstruction of missing data in both binary and continuous AMR-related datasets, under MCAR and MNAR mechanisms and for increasing levels of missingness (10–50%). For binary data, the comparative analysis on the public and synthetic datasets showed a clear and consistent hierarchy among methods. MICE emerged as the most robust approach in almost all scenarios, achieving the highest accuracy, F_1-score and AUC even at high missing rates and under MNAR missingness, where the reconstruction task is intrinsically more challenging. RF generally provided intermediate performance, remaining competitive at lower missing percentages but degrading more markedly as missingness increased, whereas KNN was systematically the least effective and the most sensitive to data sparsity. These results indicate that, when the goal is to preserve the structure of binary resistance/virulence profiles, model-based and iterative procedures such as MICE are preferable to simpler, purely distance-based imputers.

For continuous ARG abundance datasets, MICE does not always maintain the best performances, despite continuing to give good results. Across all five habitats and for both MCAR and MNAR mechanisms, KNN, KMI and MICE yielded very low NMAE and NRMSE values, with only modest differences between methods. Error levels remained remarkably stable as the proportion of missing values increased, and the ranking among methods was often dataset-dependent. MICE and KNN alternated as best performers on different dataset–metric combinations, while KMI was typically close but rarely optimal. These findings suggest that, in the presence of sufficiently rich and structured continuous data, all three approaches are capable of reconstructing quantitative ARG profiles with limited distortion, and the practical choice among them may be driven more by computational considerations and integration with downstream analyzes than by raw imputation accuracy.

Missing (%)	Method	Dataset									
		SL		WT		SD		PM		DT	
		NMAE	NRMSE	NMAE	NRMSE	NMAE	NRMSE	NMAE	NRMSE	NMAE	NRMSE
10	KNN	**0.0010**	0.0049	0.0009	**0.0076**	**0.0011**	**0.0055**	0.0800	0.1335	0.1539	0.4170
	KMI	0.0017	0.0086	0.0020	0.0081	0.0017	0.0066	0.0887	0.1527	0.2527	0.4466
	MICE	**0.0010**	**0.0046**	**0.0008**	0.0083	0.0012	0.0067	**0.0772**	**0.1264**	**0.1342**	**0.3629**
20	KNN	0.0010	0.0047	0.0012	**0.0167**	**0.0010**	**0.0046**	**0.0778**	**0.1302**	0.1528	0.4356
	KMI	0.0014	0.0083	0.0019	0.0170	0.0023	0.0062	0.0967	0.1430	0.2807	0.4928
	MICE	**0.0008**	**0.0045**	**0.0010**	0.0172	0.0013	0.0064	0.0788	0.1327	**0.1310**	**0.3388**
30	KNN	**0.0010**	0.0049	0.0011	**0.0158**	**0.0012**	**0.0069**	**0.0806**	0.1415	0.1600	0.4652
	KMI	0.0023	0.0083	**0.0014**	0.0159	0.0024	0.0082	0.1084	0.1514	0.2119	0.5319
	MICE	**0.0010**	**0.0046**	0.0010	0.0160	0.0014	0.0078	0.0821	**0.1403**	**0.1394**	**0.3773**
40	KNN	0.0011	0.0050	**0.0011**	**0.0176**	**0.0013**	**0.0070**	0.0829	0.1406	0.1664	0.4934
	KMI	0.0022	0.0083	0.0014	0.0177	0.0023	0.0080	0.0938	0.1516	0.2203	0.5392
	MICE	**0.0010**	**0.0047**	**0.0011**	0.0195	0.0015	0.0075	**0.0794**	**0.1357**	**0.1554**	**0.4238**
50	KNN	**0.0011**	0.0051	**0.0013**	0.0200	**0.0014**	**0.0071**	0.0854	0.1413	0.1779	0.5228
	KMI	0.0021	0.0083	0.0014	**0.0188**	0.0021	0.0080	0.0933	0.1532	0.2470	0.5448
	MICE	**0.0011**	**0.0049**	**0.0013**	0.0223	0.0015	0.0073	**0.0831**	**0.1399**	**0.1631**	**0.4179**

Table 7.6: Normalized MAE and RMSE as missing percentage (MCAR) varies for continuous datasets. For each subdataset, missing level and metric, the best (lowest) value across methods is in **bold**.

Missing (%)	Method	Dataset									
		SL		WT		SD		PM		DT	
		NMAE	NRMSE	NMAE	NRMSE	NMAE	NRMSE	NMAE	NRMSE	NMAE	NRMSE
10	KNN	**0.0007**	**0.0033**	**0.0009**	**0.0071**	**0.0009**	**0.0030**	0.1107	0.1764	0.2047	0.5368
	KMI	0.0017	0.0069	0.0020	0.0258	0.0017	0.0067	0.0887	0.1423	0.2527	0.5496
	MICE	0.0013	0.0053	0.0011	0.0235	0.0012	0.0053	**0.0689**	**0.1131**	**0.1755**	**0.4611**
20	KNN	**0.0007**	**0.0034**	**0.0010**	**0.0103**	**0.0013**	**0.0068**	0.0986	0.1655	0.1480	0.3829
	KMI	0.0014	0.0051	0.0019	0.0303	0.0023	0.0090	0.0967	0.1551	0.2807	0.6600
	MICE	0.0011	0.0046	0.0012	0.0199	0.0026	0.0116	**0.0871**	**0.1532**	**0.1743**	**0.4487**
30	KNN	**0.0009**	**0.0038**	0.0013	0.0193	0.0016	0.0101	0.1088	0.1905	0.1935	0.5620
	KMI	0.0023	0.0085	**0.0014**	**0.0204**	0.0024	0.0095	0.1084	0.1970	**0.2119**	**0.4978**
	MICE	0.0010	0.0042	0.0020	0.0173	**0.0017**	**0.0091**	**0.1055**	**0.1875**	0.2010	0.5442
40	KNN	**0.0010**	0.0048	0.0012	0.0234	**0.0018**	0.0090	0.0827	**0.1371**	0.1931	0.5563
	KMI	0.0022	0.0080	0.0014	0.0235	0.0023	0.0104	0.0938	0.1558	0.2203	0.4726
	MICE	0.0010	**0.0047**	**0.0011**	**0.0220**	0.0022	**0.0092**	**0.1021**	0.1719	**0.1799**	**0.4759**
50	KNN	0.0011	0.0052	**0.0012**	**0.0175**	0.0015	**0.0088**	0.0909	**0.1485**	0.1996	0.5864
	KMI	0.0021	0.0079	0.0014	0.0220	0.0021	0.0097	0.0933	0.1551	0.2470	0.5657
	MICE	**0.0010**	**0.0048**	0.0012	0.0233	**0.0017**	0.0071	**0.0877**	0.1397	**0.1792**	**0.4758**

Table 7.7: Normalized MAE and RMSE as missing percentage (MNAR) varies for continuous datasets. For each subdataset, missing level and metric, the best (lowest) value across methods is in **bold**.

Part II

Population models

Chapter 8

Introduction to Epidemic Models

Mathematical epidemic models are fundamental tools for describing and predicting the dynamics of infectious diseases. Their strength lies in their ability to translate biological processes into formal systems, enabling the analysis of equilibrium states, stability properties, and the effects of intervention strategies.

This chapter introduces the modeling frameworks that is used in the second part of the thesis, with the aim of establishing a consistent methodological basis for subsequent applications to AMR. After a brief historical overview, we present elementary compartmental models (SI, SIR, SIS), which constitute the minimum building blocks for more complex epidemiological frameworks. We then consider multi-strain extensions, in which competition between susceptible and resistant pathogens is explicitly represented, providing a direct mechanistic basis for analyses of AMR. Finally, we discuss network-based and spatial formulations, showing how metapopulation models incorporate heterogeneity in connectivity and mobility and how these characteristics influence epidemic spread.

8.1 Historical background and conceptual framework

The origins of mathematical epidemiology date back to the nineteenth century, when early attempts were made to quantify the spread of infectious diseases using demographic and statistical observations. Daniel Bernoulli's model for smallpox (1760) is often considered the first analytical approach to the dynamics of infection within a population, followed later by the works of Ross on malaria (1911) and Hamer on measles (1906) [47]. These pioneering studies laid the conceptual foundation for linking infection processes with population-level dynamics, even if the models were still relatively simple and lacked a general theoretical framework.

A decisive step forward was achieved with the work of Kermack and McKendrick (1927) [53], who formulated the first general mathematical model for epidemic propagation in a homogeneous population. Their system of differential equations, known as the SIR model, introduced the idea that an epidemic ends not because all susceptible individuals are infected, but because the susceptible fraction falls below a critical threshold. This threshold concept, together with the introduction of the basic reproduction number R_0, became the cornerstone of modern epidemiological modeling [4, 47].

Subsequent decades witnessed a steady evolution of modeling approaches. Deterministic compartmental models were extended to include demographic processes, latency periods, waning immunity, and spatial or network structures. Stochastic formulations were developed to describe the inherent randomness of infection events in finite populations. The metapopulation framework emerged to account for heterogeneous mixing and spatial coupling among subpopulations, while agent-based models provided an explicit representation of individual behaviors and contact networks. Brauer (2017) [15] provides a comprehensive overview of this historical trajectory, highlighting how advances in computational power and data availability have continuously expanded the scope and realism of epidemic models.

The COVID-19 pandemic further underscored the importance of mathematical epidemiology as a decision-support tool. Models, ranging from simple compartmental structures to high-resolution stochastic simulations,

played a central role in estimating transmission parameters, evaluating intervention scenarios, and informing public health policies [111, 44]. At the same time, the crisis exposed the limitations of traditional homogeneous-mixing models, highlighting the need for more flexible frameworks that incorporate heterogeneity, mobility, and behavioural feedback. Network and mobility-based studies provided key evidence of how spatial structure and travel patterns shaped epidemic trajectories [19, 40, 25].

8.2 Examples of standard models and approaches to disease spreading

Mathematical epidemiology provides a broad family of models aimed at describing how infectious diseases spread through populations. Despite the diversity of approaches, most models share the same fundamental structure: they represent the movement of individuals between compartments corresponding to distinct epidemiological states, such as susceptible, infectious, recovered, or exposed. Depending on the level of description and the underlying assumptions, these models can be formulated either at the population level or at the level of individual entities.

8.2.1 Classical compartmental models

The classical compartmental framework includes a series of well-established models that describe the temporal evolution of epidemics in a homogeneous population. The simplest one is the *SIS* model, where individuals move from the susceptible state to the infectious one and return to susceptibility after recovery. This structure is suitable for diseases that do not confer immunity, such as certain bacterial infections.

The *SIR* model, introduced by Kermack and McKendrick (1927) [53], adds a removed (or recovered) compartment, representing individuals who have acquired immunity or are no longer infectious. It has become the archetype of deterministic epidemic models and remains the foundation for numerous generalizations. The *SEIR* model extends this structure by adding an *exposed* compartment, accounting for a latent period between infection and infectiousness. Other variants, such as SIRS, SEIRS, or models incor-

porating births, deaths, and vaccination, have been developed to capture different biological and demographic features [47, 15].

Although these models are often presented in deterministic, continuous-time form, they can also be formulated as discrete-time systems or as stochastic processes to reflect random effects in small populations. Their analytical tractability makes them useful for studying epidemic thresholds, equilibrium states, and the effects of interventions such as vaccination or social distancing.

8.2.2 Modeling disease spreading: population vs. individual level

The previous section introduced population-level models, in which disease dynamics are described through aggregated compartments. These models assume homogeneous mixing within each group and are typically formulated as systems of differential or difference equations. Their analytical tractability enables the derivation of global epidemiological indicators, such as the basic reproduction number R_0, epidemic thresholds, and the expected impact of control measures, and includes both single-population and metapopulation frameworks, the latter representing interacting subpopulations coupled through mobility flows [4, 52].

When the assumption of homogeneous mixing is insufficient, a finer level of description becomes necessary. Network-based models represent individuals as nodes of a contact graph, where edges encode potentially infectious interactions. In these settings, epidemic dynamics depend on the topology of the network and on the statistical distribution of contacts [85]. A further refinement is provided by agent-based models (ABMs), where each individual is described explicitly together with behavioural rules, mobility patterns, and stochastic interactions. These models offer a microscopic and data-driven representation of epidemic spread and have become particularly useful for simulating complex scenarios such as those observed during the COVID-19 pandemic [52, 15].

8.3 Metapopulation theory

Classical compartmental models assume a well-mixed population, where every individual has the same probability of contacting any other. In real systems, however, populations are often structured in space or divided into distinct subgroups (such as cities, hospitals, farms, or social clusters) connected through mobility or contact networks. The *metapopulation framework* was introduced to capture this heterogeneity and to describe the coupled dynamics of multiple interacting populations [42, 68, 52].

In a metapopulation model, each subpopulation (or *patch*) hosts its own local epidemic dynamics, typically described by a compartmental model such as SIS, SIR, or SEIR. Individuals (or pathogens) can move between patches according to a connectivity structure, which may be represented by a mobility matrix or a weighted network. This coupling enables the study of how local outbreaks can trigger regional or global epidemics and how spatial fragmentation or limited mobility can slow or prevent disease invasion [23, 52].

This framework allows us to address a wide range of questions concerning the spatial and structural dynamics of epidemics. In particular, it makes it possible to determine under which conditions an infection can successfully invade and persist across a network of interconnected subpopulations. It also provides a means to evaluate how local interventions, such as vaccination campaigns, quarantine policies, or travel restrictions, may alter the global progression of the disease. Moreover, the metapopulation approach highlights the role of heterogeneity: differences in population size, connectivity, or mobility intensity can profoundly influence the epidemic threshold and the long-term persistence of infection.

Metapopulation theory has been successfully applied to a variety of contexts, from childhood diseases in spatially structured populations [42] to the modelling of global pandemic spread through air-transportation networks [23].

Chapter 9

Theoretical framework: reproduction numbers and Next Generation Matrix

9.1 Introduction

A central concept in mathematical epidemiology is the *basic reproduction number*, commonly denoted as R_0. It represents the expected number of secondary infections generated by a single infectious individual introduced into a fully susceptible population [4, 27]. This indicator provides a threshold criterion for the potential spread of an infectious disease: if $R_0 > 1$, the infection can invade the population and possibly become endemic; conversely, if $R_0 < 1$, the infection is expected to die out. Therefore, R_0 serves as a fundamental parameter for assessing the stability of the disease-free equilibrium and for designing and evaluating control strategies.

From a biological point of view, R_0 quantifies the average transmission potential of an infection, integrating the combined effects of contact rate, transmission probability, and duration of the infectious period. Despite its apparent simplicity, the computation and interpretation of R_0 depend heavily on the structure of the underlying epidemiological model. In homogeneous models such as SIS or SIR, R_0 can be derived analytically as a simple ratio between transmission and recovery rates. However, in more realistic settings, where heterogeneities in host population, contact patterns,

or spatial structure are considered, the estimation of R_0 requires a more general and formal approach. To address this, the *Next Generation Matrix* (NGM) framework was developed [110, 29]. This formalism provides a systematic method to compute reproduction numbers in models that include multiple infectious compartments or interacting subpopulations. Within this framework, the infection process is decomposed into two components: the generation of new infections and the transitions of individuals between infected classes. The resulting matrix representation allows R_0 to be expressed as the spectral radius (dominant eigenvalue) of the next generation matrix, offering a mathematically rigorous and biologically interpretable definition.

In addition to R_0, extended indicators have been proposed to capture the invasion potential of a pathogen in structured systems. Among them, the *invasion reproduction number* (R_{inv}) measures the ability of an infection or a new strain to invade a population where another infection is already present [45, 28]. This concept generalizes the idea of a threshold parameter, playing a similar role in determining whether an invading pathogen can establish itself.

The following sections introduce the theoretical background necessary to derive R_0 and related indicators. We first recall the formulation of simple compartmental models such as SIS and SIR, where R_0 can be obtained directly. We then extend the discussion to more complex systems, where the Next Generation Matrix formalism provides a general and unifying theoretical framework for computing reproduction numbers across a wide class of epidemic models.

9.2 Calculation of R_0 in simple models

The concept of the basic reproduction number can be first illustrated through simple compartmental models, which provide an intuitive understanding of how transmission and recovery processes determine whether an infection can persist in a population. In this section, we consider two classical models: the SIS and the SIR models. These simplified formulations serve as a starting point for the generalization introduced later through the NGM framework.

9.2.1 The SIS model

In the SIS (Susceptible–Infected–Susceptible) model, individuals move from the susceptible class (S) to the infected class (I) upon contact with an infected individual, and then return to the susceptible class after recovery. No immunity is assumed, so recovered individuals are immediately susceptible again. The total population size N is constant, and $S + I = N$.

If we consider the fraction of population in each compartment ρ, such that $\rho^S + \rho^I = 1$, the model dynamics are described by

$$\begin{cases} \frac{d\rho^S}{dt} = -\beta\rho^S\rho^I + \mu\rho^I, \\ \frac{d\rho^I}{dt} = \beta\rho^S\rho^I - \mu\rho^I, \end{cases} \tag{9.1}$$

where β is the transmission rate and μ the recovery rate.

At the disease-free equilibrium, where $\rho^I \approx 0$, $\rho^S \approx 1$, the stability of the equilibrium is assessed by linearising the second equation around this point:

$$\frac{d\rho^I}{dt} \approx (\beta - \mu)\rho^I.$$

If the linearised system has a positive growth rate, i.e. if $\beta > \mu$, the disease-free equilibrium is unstable and an outbreak can occur. This motivates the definition of the basic reproduction number:

$$R_0 = \frac{\beta}{\mu}, \tag{9.2}$$

which separates two regimes: when $R_0 > 1$, the system admits a stable endemic equilibrium; when $R_0 < 1$, the disease-free equilibrium is stable, and the infection cannot persist.

9.2.2 The SIR model

In the SIR (Susceptible–Infected–Removed) model, individuals who recover move into a distinct compartment (R) and acquire permanent immunity. lso in this case we can consider the fraction of population: $\rho^S + \rho^I + \rho^R = 1$. The model equations are:

$$\begin{cases} \frac{d\rho^S}{dt} = -\beta\rho^S\rho^I, \\ \frac{d\rho^I}{dt} = \beta\rho^S\rho^I - \mu\rho^I, \\ \frac{d\rho^R}{dt} = \mu\rho^I. \end{cases} \tag{9.3}$$

with the same definitions of β and μ as in the SIS case.

At the beginning of the epidemic, $\rho^S \approx 1$, so the dynamics of the infected class are well approximated by the linearised form:

$$\frac{d\rho^I}{dt} \approx (\beta - \mu)\rho^I.$$

As in the SIS case, the infection initially grows if and only if $\beta > \mu$, which leads to the same threshold expression

$$R_0 = \frac{\beta}{\mu}.$$

Although R_0 has the same algebraic form as in the SIS model, the long-term behaviour differs: in the SIR framework, immunity accumulates, reducing the susceptible fraction and thus the effective transmission rate. The epidemic peaks when $\rho^S < 1/R_0$ and eventually dies out as immunity increases.

Even in these simple models, R_0 acts as a threshold parameter capturing the combined effect of transmission, recovery, and population structure, providing a unified criterion for disease invasion and control.

9.3 Extension to more complex models

While simple compartmental models such as SIS and SIR provide fundamental insights into infection dynamics, they rely on strong simplifying assumptions, most notably the homogeneity of the population. In real systems, transmission processes are influenced by heterogeneities in contact patterns, demographic structure, spatial distribution, and biological characteristics of the host and pathogen. These heterogeneities can profoundly affect the expression of the basic reproduction number R_0.

9.3.1 Incorporating additional compartments

A natural way to refine compartmental models is to introduce additional compartments that represent specific stages of the infection process. This refinement can capture biologically relevant features but comes at the cost of increasing the number of parameters, which may limit identifiability and, in some cases, reduce the practical realism of the model. One of the most

common extensions is the SEIR model, which includes an *exposed* class (E) representing individuals who have been infected but are not yet infectious. The model equations are typically written as:

$$\begin{cases} \frac{d\rho^S}{dt} = -\beta\rho^S\rho^I, \\ \frac{d\rho^E}{dt} = \beta\rho^S\rho^I - \sigma\rho^E, \\ \frac{d\rho^I}{dt} = \sigma\rho^E - \mu\rho^I, \\ \frac{dR}{dt} = \mu\rho^I, \end{cases} \tag{9.4}$$

where σ^{-1} represents the average latency period before an individual becomes infectious.

The presence of multiple infectious or partially infectious compartments (e.g., asymptomatic, symptomatic, hospitalized) complicates the analytical derivation of R_0, since the contribution of each class to the generation of new infections must be accounted for explicitly. In such cases, the intuitive ratio β/μ used in simple models is no longer valid, and a more general formalism is required.

9.3.2 Heterogeneity in population structure

In many epidemiological contexts, the assumption of a homogeneous population is unrealistic. Differences in age, social behavior, or spatial distribution often generate structured mixing patterns that significantly influence the transmission dynamics of infectious diseases. In *age-structured models*, for instance, the rate of infection depends on empirically derived contact matrices that quantify interactions between individuals of different age groups [4]. These matrices capture heterogeneities in social behavior, such as school attendance or occupational exposure, which play a crucial role in determining transmission intensity across age classes.

Similarly, in *metapopulation models*, the host population is subdivided into distinct subpopulations or patches, connected through processes such as mobility, commuting, or migration [21]. In these systems, the spread of an infection depends not only on local epidemiological parameters but also on the connectivity structure linking different subpopulations, which determines the rate and direction of pathogen dispersal.

A further layer of complexity arises in *multi-strain models*, where multiple pathogen variants coexist and compete within the same host population [28]. In such cases, the invasion potential of a new strain depends on the cross-immunity conferred by existing infections and on the ecological interactions among competing strains. Altogether, these structured approaches demonstrate how heterogeneity in contact patterns, spatial organization, and pathogen diversity fundamentally shapes epidemic dynamics and challenges the straightforward computation of R_0.

Such structured models require a systematic method to quantify how infections are generated and transmitted across different compartments or subpopulations. Linearizing these models around the disease-free equilibrium often leads to systems of equations that can be expressed in matrix form, where each element represents the expected number of secondary infections produced by one individual in a given compartment.

9.3.3 Towards a general formalism

To handle this complexity, a unified mathematical framework is needed to compute the reproduction number across diverse model structures. This need motivated the development of the NGM approach [29, 110, 28]. The NGM formalism generalizes the concept of R_0 to models with multiple infectious compartments or heterogeneous mixing by expressing transmission and transition processes in matrix form.

In the next section, we introduce the theoretical foundation of the NGM, showing how it provides a rigorous and general definition of the basic reproduction number as the dominant eigenvalue of a matrix that encapsulates the generation of new infections. This approach not only extends the applicability of R_0 but also offers a clear link between epidemiological mechanisms and the dynamical stability of epidemic models.

9.4 The Next Generation Matrix

The NGM formalism provides a rigorous and general framework to compute the basic reproduction number in epidemiological models that include multiple infectious compartments or interacting subpopulations. It was originally introduced by Diekmann, Heesterbeek, and Metz [29], and later formalized

and generalized by van den Driessche and Watmough [110]. The method is based on linearizing the infection dynamics around the disease-free equilibrium and quantifying the expected number of secondary infections generated by each infectious compartment.

9.4.1 Conceptual basis

The central idea behind the NGM approach is to decompose the system of differential (or difference) equations describing epidemic dynamics into two distinct components: the *new infection terms*, which represent the appearance of newly infected individuals in each compartment, and the *transition terms*, which describe changes among existing infected classes, including progression, recovery, or removal. By isolating these two processes, the method allows the infection subsystem to be represented as a set of coupled linear equations that capture how infections propagate from one class to another near the disease-free equilibrium.

Consider a model with n infectious compartments, collected in the vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. In continuous time, the infection dynamics can generally be expressed as:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}) - \mathbf{V}(\mathbf{x}), \tag{9.5}$$

where $\mathbf{F}(\mathbf{x})$ is the vector of rates at which new infections appear in each compartment, and $\mathbf{V}(\mathbf{x})$ is the vector of rates describing transitions between infectious compartments or exits due to recovery or death.

The Jacobian matrices of $\mathbf{F}(\mathbf{x})$ and $\mathbf{V}(\mathbf{x})$, evaluated at the disease-free equilibrium $\mathbf{x} = \mathbf{0}$, are denoted as F and V, respectively:

$$F = \left[\frac{\partial F_i}{\partial x_j}(\mathbf{0})\right], \qquad V = \left[\frac{\partial V_i}{\partial x_j}(\mathbf{0})\right]. \tag{9.6}$$

The NGM is then defined as:

$$K = FV^{-1}. \tag{9.7}$$

Each element K_{ij} represents the expected number of new infections in compartment i caused by one individual initially introduced into compartment j, during its entire infectious period. The basic reproduction number is given by the spectral radius of this matrix:

$$R_0 = \rho(K) = \rho(FV^{-1}), \tag{9.8}$$

where $\rho(\cdot)$ denotes the dominant eigenvalue. The disease-free equilibrium is stable if $\rho(FV^{-1}) < 1$ and unstable otherwise.

The NGM theory can be equivalently formulated for systems evolving in discrete time, which are often used when the infection process is represented through agent-based models [28]. Let $\mathbf{X}(t)$ denote the vector of the infectious compartments at time t. The evolution of the system near the disease-free equilibrium can be linearized as:

$$\mathbf{X}(t+1) = (T + \Sigma)\mathbf{X}(t), \tag{9.9}$$

where T is the *Transmission Matrix*, describing the appearance of new infections, and Σ is the *Transition Matrix*, describing movements among existing infected classes and recoveries.

Analogously to the continuous case, the NGM in discrete time is obtained by:

$$K = T(\mathcal{I} - \Sigma)^{-1}, \tag{9.10}$$

where $\mathcal{I}$ is the identity matrix. The term $(\mathcal{I}-\Sigma)^{-1}$ represents the cumulative effect of successive transitions among infectious classes over discrete time steps. The basic reproduction number is then computed as the spectral radius of K:

$$R_0 = \rho(K) = \rho\left[T(\mathcal{I} - \Sigma)^{-1}\right]. \tag{9.11}$$

This discrete-time formulation is fully consistent with the continuous-time version, and the two coincide in the limit of infinitesimally small time steps. The choice between the two approaches depends on the temporal structure of the model and on whether the underlying data or processes are better represented as continuous flows or as discrete generations of infection.

9.5 Extensions and variations

The NGM formalism not only provides a rigorous definition of the basic reproduction number R_0, but it also enables the derivation of other related indicators that describe more complex epidemic situations. Among these, the most relevant generalization is the *invasion reproduction number*, denoted as R_{inv}, which quantifies the ability of a pathogen (or a strain) to invade a population that is not entirely susceptible.

9.5.1 The invasion reproduction number

The concept of R_{inv} arises naturally in models describing multiple host populations, multiple strains of a pathogen, or multiple infectious agents competing within the same system [45, 28]. While R_0 is computed around the disease-free equilibrium, R_{inv} is instead evaluated around an *endemic equilibrium* in which one or more infections are already present.

Formally, the procedure parallels that used for R_0. Let the system be described by a set of equations representing the dynamics of several interacting infections or populations. Assuming that one infection (or strain) is endemic and another is initially rare, the next-generation approach can be applied to the invading component. In continuous-time models, the corresponding Jacobian matrices F and V for the invading subsystem are evaluated at the endemic equilibrium of the resident system, yielding:

$$R_{\text{inv}} = \rho(FV^{-1})\Big|_{\text{endemic}}. \tag{9.12}$$

For discrete-time systems, an analogous formulation can be written in terms of the transmission and transition matrices, T and Σ, introduced in Eq. (9.9). By linearizing the invading subsystem around the endemic equilibrium, one obtains:

$$R_{\text{inv}} = \rho\left[T(\mathcal{I} - \Sigma)^{-1}\right]_{\text{endemic}}. \tag{9.13}$$

In both cases, $\rho(\cdot)$ denotes the spectral radius, and the subscript "endemic" indicates that the quantities are computed with respect to the equilibrium conditions of the pre-existing infection.

In the case of pure competition, that is, if each strain competes for the same susceptible population and there are no direct interactions, R_{inv} can be determined simply from the two basic reproduction numbers of the two infectious strains:

$$R_{(j|i)} = \frac{R_{0,j}}{R_{0,i}} \tag{9.14}$$

where i is the resident strain and j is the invader.

Biologically, R_{inv} measures whether the invading pathogen can increase when introduced in small numbers into a system already hosting a resident strain. If $R_{\text{inv}} > 1$, the invader can successfully spread and potentially

replace or coexist with the resident infection; if $R_{\text{inv}} < 1$, invasion fails and the resident strain remains dominant. This threshold concept plays a central role in evolutionary epidemiology and in the analysis of competition and coexistence among pathogens.

9.5.2 Applications to structured and metapopulation models

In metapopulation frameworks, R_{inv} can describe the ability of a disease to spread from a newly infected subpopulation into a system where other subpopulations are already endemic. The NGM structure (expressed either as $K = FV^{-1}$ in continuous time or $K = T(\mathcal{I}-\Sigma)^{-1}$ in discrete time) allows us to compute this value directly from the connectivity matrix describing movement between patches and from local transmission parameters [21]. Similarly, in multi-host or vector-borne disease models, R_{inv} quantifies the potential for a pathogen to invade a community where other host species or competing parasites are already circulating.

In metapopulation models with multiple strains, it is often convenient to express R_{inv} in a compact form that explicitly separates the effect of the susceptible pool from the structure of the NGM. Consider two strains, indexed by i (resident) and j (invader), spreading over a network of N patches. Let K_i and K_j denote the NGMs of the two strains computed at the disease-free equilibrium (i.e. in the absence of any other infection), and let

$$\mathbf{S}_i^* = (S_{i,1}^*, \ldots, S_{i,N}^*)^T$$

be the vector of susceptible fractions at the endemic equilibrium of strain i. We define the diagonal matrix

$$D(\mathbf{S}_i^*) = \operatorname{diag}(S_{i,1}^*, \ldots, S_{i,N}^*).$$

Linearising the equations of the invading strain j around the endemic equilibrium of strain i, the invasion reproduction number takes the form

$$R_{(j|i)} = \rho\big(D(\mathbf{S}_i^*)\, K_j\big) \tag{9.15}$$

In this formulation:

- K_j is the NGM of strain j at the disease-free equilibrium, capturing the intrinsic transmission and mobility structure of strain j in a fully susceptible metapopulation;

- $D(\mathbf{S}_i^*)$ rescales the NGM by the fraction of susceptibles remaining at the endemic equilibrium of strain i, thus encoding the depletion of susceptibles induced by the resident infection.

The condition $R_{(j|i)} > 1$ implies that strain j can increase when rare in a system where strain i is already endemic, and therefore can invade; conversely, $R_{(j|i)} < 1$ implies failure of invasion and persistence of the resident strain.

9.5.3 Relationship between R_0 and $R_{\mathbf{inv}}$

Although R_0 and R_{inv} share the same mathematical foundation, their interpretations differ. The basic reproduction number R_0 characterizes the potential for disease emergence in a fully susceptible population, whereas R_{inv} evaluates the ability of a new infection to establish itself in an already structured or partially immune environment. In practice, R_0 acts as a global threshold for the onset of epidemics, while R_{inv} provides a local threshold determining the outcome of invasion or competition processes.

From a theoretical standpoint, both indicators stem from the same spectral condition ($\rho(FV^{-1}) > 1$ in continuous time or $\rho[T(\mathcal{I} - \Sigma)^{-1}] > 1$ in discrete time) but differ in the equilibrium around which the system is linearized. This highlights the flexibility of the NGM framework, which accommodates a broad range of epidemiological contexts while maintaining mathematical consistency and biological interpretability.

The concepts developed in this chapter form the foundation for the analytical and computational tools presented in the following sections, where the theoretical framework of reproduction numbers is applied to more realistic epidemic models and quantitative simulations.

Chapter 10

A model for AMR in the hospital setting: the XSR model

In recent decades, the integration of metapopulation structures into epidemiological models has represented a substantial methodological advance. As discussed in Section 8.3, these models partition the population into interconnected subgroups, such as geographic regions, hospital wards, or patient pathways, thus capturing the impact of heterogeneous contact structures and mobility-driven transmission dynamics. This framework has proven particularly effective for healthcare-associated pathogens, whose spread is strongly shaped by patient transfers, ward-specific practices, and localized infection control interventions [22, 52].

Parallel to this, the study of interactions among multiple pathogen types has gained increasing relevance. Empirical evidence shows that individuals may carry or acquire distinct infectious agents, either sequentially or concurrently, and that competition or facilitation between pathogens can alter transmission patterns in ways that single-pathogen models fail to predict [62]. Despite advances in both metapopulation and multipathogen modeling, their integration remains limited, especially in hospital settings where spatial heterogeneity, mobility constraints, and structured patient pathways are intrinsic.

AMR further complicates this landscape. Resistant and sensitive strains

differ not only in transmissibility (often modulated by fitness costs) but also in the selective pressures imposed by antibiotic use and by ward-level infection control strategies [5]. Hospitals commonly enforce differentiated isolation, cohorting, or transfer policies depending on colonization status, generating mobility patterns that are inherently strain-dependent. A realistic model must therefore account simultaneously for strain competition, selective pressures, the possibility of transitions from sensitivity to resistance, and structured patient movement.

In this chapter, we introduce the *XSR* model, a metapopulation framework specifically designed to capture the transmission and competition between antibiotic-sensitive and antibiotic-resistant strains in a spatially structured hospital environment. The model couples compartmental epidemiological dynamics with ward-level mobility and strain-dependent transitions, providing a theoretical basis for the analyses presented in the following chapter.

10.1 Description of the model

We consider a closed population of hospital patients, whose infection status with respect to a bacterial species is described by three epidemiological classes:

- X: uninfected (or uncolonized) patients;
- S: patients infected with a strain that is susceptible to antibiotic A_1;
- R: patients infected with a strain that is resistant to antibiotic A_1.

In both infected classes (S and R), the pathogen is assumed to remain susceptible to a second antibiotic A_2. The model thus focuses on resistance to a single antimicrobial agent, embedded in a treatment context where a second agent remains effective.

Patients in S and R may recover either spontaneously or under antibiotic treatment. We denote by γ the spontaneous recovery probability, and by τ_1 and τ_2 the recovery probabilities due to administration of A_1 and A_2, respectively. The acquisition of resistance in previously susceptible infections is modeled by a transition probability σ from S to R, typically small.

Susceptible patients in class X may become infected through contact with infected individuals. We denote by β_1 the infection probability per time step for acquisition of the susceptible strain, and by β_2 the corresponding probability for the resistant strain. Fitness costs associated with resistance are incorporated by setting

$$\beta_2 = \beta_1(1-c), \qquad c \in [0,1], \tag{10.1}$$

where c is the fitness cost of resistance [93, 6]. A larger c implies a more pronounced reduction in the transmissibility of the resistant strain.

The compartmental structure of the XSR model is illustrated in Fig. 10.1. Patients can move from X to S or R via infection, from S to R via acquisition of resistance, and from both infected classes to X via spontaneous or treatment-induced recovery.

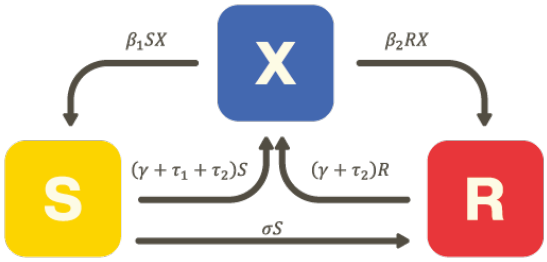


Figure 10.1: **XSR model:** patients can be healthy (X), infected by a susceptible strain (S) or infected by a resistant strain (R). Parameters β_1 and β_2 denote the infection probabilities for the susceptible and resistant strains, respectively; γ is the spontaneous recovery probability; τ_1 and τ_2 are the recovery probabilities under antibiotics A_1 and A_2; σ is the probability that a susceptible infection becomes resistant (typically very small).

10.1.1 Single-ward dynamics and stability analysis

We first consider the dynamics in a single, well-mixed ward, ignoring mobility. Let $S(t)$ and $R(t)$ denote, respectively, the fractions of patients in classes S and R at discrete time t; the fraction of susceptible (uninfected) patients is then:

$$X(t) = 1 - S(t) - R(t). \tag{10.2}$$

The discrete-time evolution equations for the infected compartments are:

$$\begin{aligned} S(t+1) &= S(t) + \beta_1 S(t)X(t) - (\gamma + \tau_1 + \tau_2)S(t) - \sigma S(t), \\ R(t+1) &= R(t) + \beta_2 R(t)X(t) + \sigma S(t) - (\gamma + \tau_2)R(t). \end{aligned} \tag{10.3}$$

The terms proportional to β_1 and β_2 describe infection events, those proportional to γ, τ_1, τ_2 describe recovery, and the term proportional to σ describes acquisition of resistance.

Imposing stationarity conditions $S(t+1) = S(t) = S^*$ and $R(t+1) = R(t) = R^*$ in (10.3), and using $X^* = 1 - S^* - R^*$, we obtain:

$$\begin{aligned} S^* \left[\beta_1(1 - S^* - R^*) - (\gamma + \tau_1 + \tau_2 + \sigma)\right] &= 0, \\ R^* \left[\beta_2(1 - S^* - R^*) - (\gamma + \tau_2)\right] + \sigma S^* &= 0. \end{aligned} \tag{10.4}$$

From (10.4) we identify four fixed points, corresponding to different epidemiological regimes:

1. Disease-free equilibrium:

$$(S^*, R^*) = (0, 0).$$

2. Endemic equilibrium with only the susceptible strain:

$$(S^*, R^*) = \left(1 - \frac{\gamma + \tau_1 + \tau_2 + \sigma}{\beta_1},\ 0\right),$$

   provided that $\beta_1 > \gamma + \tau_1 + \tau_2 + \sigma$.

3. Endemic equilibrium with only the resistant strain:

$$(S^*, R^*) = \left(0,\ 1 - \frac{\gamma + \tau_2}{\beta_2}\right),$$

   provided that $\beta_2 > \gamma + \tau_2$.

4. Endemic equilibrium with coexistence of both strains:

   (S^*, R^*) such that $S^* \neq 0$ and $R^* \neq 0$. The explicit closed-form expressions are algebraically cumbersome and are therefore reported in Appendix A.1. Here, we focus on the conditions for the existence of such a coexistence equilibrium.

In practice, the parameter σ is often sufficiently small that it can be approximated by zero. Under this assumption, the coexistence equilibrium ($S^* \neq 0$, $R^* \neq 0$) exists only in the special case where:

$$\frac{\gamma + \tau_1 + \tau_2}{\beta_1} = \frac{\gamma + \tau_2}{\beta_2}, \tag{10.5}$$

i.e. when the effective reproduction ratios of the two strains coincide. In all other parameter regimes, the two infected states are in competition, and one of the two strains eventually dominates.

To study the local stability of the equilibria, we compute the Jacobian matrix of system (10.3) with respect to (S, R):

$$J(S,R) = \begin{bmatrix} -2\beta_1 S + [1 + \beta_1 - (\gamma + \tau_1 + \tau_2 + \sigma)] - \beta_1 R & -\beta_1 S \\ -\beta_2 R + \sigma & -2\beta_2 R + [1 + \beta_2 - (\gamma + \tau_2)] - \beta_2 S \end{bmatrix}. \tag{10.6}$$

For a discrete-time system, a fixed point is locally stable if all eigenvalues of the Jacobian evaluated at that point have modulus strictly less than one.

As an example, consider the disease-free equilibrium $(S^*, R^*) = (0, 0)$. The Jacobian becomes:

$$J(0,0) = \begin{bmatrix} 1 + \beta_1 - (\gamma + \tau_1 + \tau_2 + \sigma) & 0 \\ \sigma & 1 + \beta_2 - (\gamma + \tau_2) \end{bmatrix}. \tag{10.7}$$

The corresponding eigenvalues are:

$$\begin{aligned} \lambda_1 &= 1 + \beta_1 - (\gamma + \tau_1 + \tau_2 + \sigma), \\ \lambda_2 &= 1 + \beta_2 - (\gamma + \tau_2). \end{aligned} \tag{10.8}$$

The disease-free equilibrium is locally stable if $|\lambda_1| < 1$ and $|\lambda_2| < 1$. In particular, the conditions:

$$\beta_1 < \gamma + \tau_1 + \tau_2 + \sigma, \qquad \beta_2 = \beta_1(1 - c) < \gamma + \tau_2 \tag{10.9}$$

are sufficient for stability. Equivalently, the disease-free state is stable if:

$$\beta_1 < \min\left\{\gamma + \tau_1 + \tau_2 + \sigma,\ \frac{\gamma + \tau_2}{1 - c}\right\}. \tag{10.10}$$

From a biological point of view, these conditions express an intuitive balance between transmission and recovery: if infection rates are sufficiently low relative to the combined effects of spontaneous recovery and treatment, then both strains die out, and the system converges to the disease-free equilibrium. Conversely, when these inequalities are violated, one or both strains can persist endemically.

Figure 10.2 illustrates the four qualitative dynamical regimes corresponding to the four equilibrium scenarios described above.

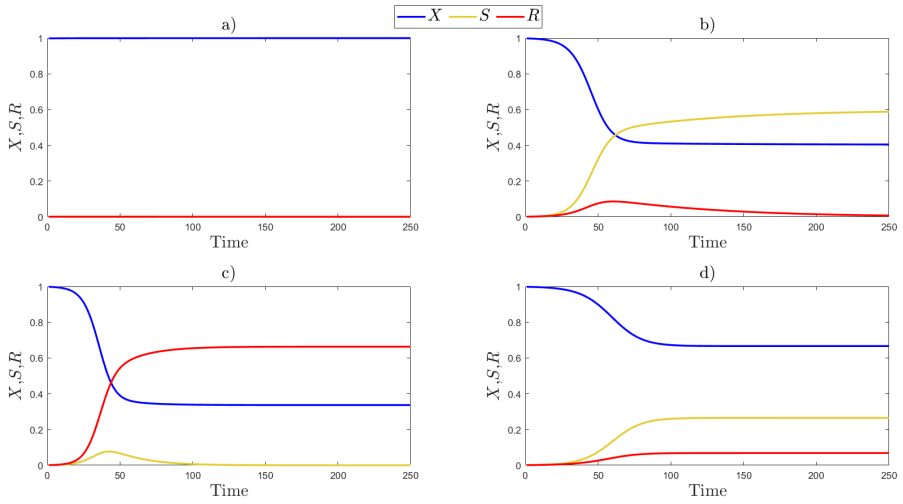


Figure 10.2: **System behaviour in the four equilibrium regimes**: (a) disease-free equilibrium, both S and R go to zero; (b) only the susceptible strain persists; (c) only the resistant strain persists; (d) coexistence of susceptible and resistant infections.

10.2 Metapopulation implementation

We now extend the XSR model to a metapopulation framework in which the hospital is represented as a network of N wards. Each ward $i = 1, \ldots, N$ is a node in the network, hosting a subpopulation of patients, and edges represent admissible patient transfers between wards.

We denote by $\rho_i^S(t)$ and $\rho_i^R(t)$ the fractions of patients in ward i at time t who are infected with the susceptible and resistant strains, respectively.

The fraction of uninfected patients in ward i is then:

$$\rho_i^X(t) = 1 - \rho_i^S(t) - \rho_i^R(t). \tag{10.11}$$

The hospital mobility structure is encoded in a weighted adjacency matrix $W = (W_{ij})$, where W_{ij} represents the weight of the connection from ward i to ward j. Patient movements are governed by a diffusion probability $p_d \in [0, 1]$.

We adopt the following notation:

- N: number of wards (nodes);
- $\rho_i^S(t), \rho_i^R(t), \rho_i^X(t)$: fractions of patients in S, R, and X in node i at time t;
- n_i: population size associated with node i;
- $n_{j\to i}$: number of patients present in node i at time t who belong to node j (including those who did not move when $i = j$);
- $n_i^{\text{eff}} = \sum_j n_{j\to i}$: effective population in node i after movement;
- W_{ij}: weight of the connection from node i to node j;
- $R_{ij} = \dfrac{W_{ij}}{\sum_l W_{il}}$: row-stochastic mobility matrix;
- p_d: probability that a patient moves from its home node to a neighbouring node.

The evolution equations for the infected fractions in node i generalize system (10.3) and are given by:

$$\rho_i^S(t+1) = [1 - (\gamma + \tau_1 + \tau_2 + \sigma)]\, \rho_i^S(t) + \left[1 - \rho_i^S(t) - \rho_i^R(t)\right] \Pi_i^{X\to S}(t),$$

$$\rho_i^R(t+1) = [1 - (\gamma + \tau_2)]\, \rho_i^R(t) + \sigma \rho_i^S(t) + \left[1 - \rho_i^S(t) - \rho_i^R(t)\right] \Pi_i^{X\to R}(t), \tag{10.12}$$

where $\Pi_i^{X\to S}(t)$ and $\Pi_i^{X\to R}(t)$ denote the probabilities that a susceptible patient associated with node i becomes infected with the susceptible or resistant strain, respectively, during the time interval $[t, t+1]$.

The recovery parameters must satisfy:

$$\gamma + \tau_1 + \tau_2 + \sigma \leq 1, \qquad \gamma + \tau_2 \leq 1, \tag{10.13}$$

to ensure that the fractions remain well-defined (no more patients recover than those present in the corresponding compartment).

The infection probabilities are given by:

$$\begin{aligned} \Pi_i^{X\to S}(t) &= (1-p_d)P_i^S(t) + p_d \sum_{j=1}^{N} \frac{W_{ij}}{\sum_{l=1}^{N} W_{il}} P_j^S(t), \\ \Pi_i^{X\to R}(t) &= (1-p_d)P_i^R(t) + p_d \sum_{j=1}^{N} \frac{W_{ij}}{\sum_{l=1}^{N} W_{il}} P_j^R(t), \end{aligned} \tag{10.14}$$

where $P_i^S(t)$ and $P_i^R(t)$ are the probabilities that a susceptible patient present in node i at time t becomes infected with the susceptible or resistant strain due to contacts with infected patients in the same node. These probabilities account for contributions from all origins j:

$$\begin{aligned} P_i^S(t) &= 1 - \prod_{j=1}^{N} \left[1 - \beta_1 \rho_j^S(t)\right]^{n_{j\to i}}, \\ P_i^R(t) &= 1 - \prod_{j=1}^{N} \left[1 - \beta_2 \rho_j^R(t)\right]^{n_{j\to i}}. \end{aligned} \tag{10.15}$$

The quantities $n_{j\to i}$ are given by:

$$n_{j\to i} = \delta_{ij}(1-p_d)\, n_i + p_d \frac{W_{ji}}{\sum_{l=1}^{N} W_{jl}}\, n_j, \tag{10.16}$$

where δ_{ij} is the Kronecker delta. For $i = j$, $n_{j\to i}$ includes those patients who remain in their home node; for $i \neq j$, it counts patients who move from node j to node i.

Equations (10.12)–(10.16) define the full metapopulation XSR model. A schematic representation of the hospital network interpretation is shown in Fig. 10.3.

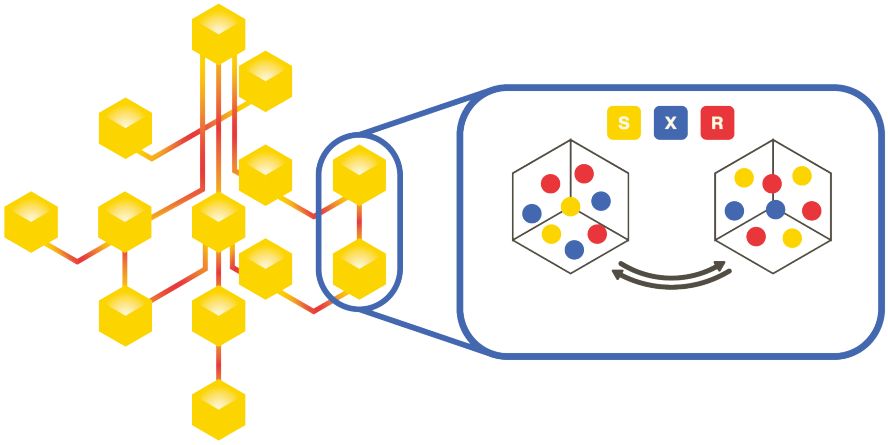


Figure 10.3: **Representation of the metapopulation model in a hospital setting:** on the left, hospital wards and their connections are shown. Within each ward (yellow cubes), patients are partitioned into three compartments: healthy (X, blue), infected with the susceptible strain (S, yellow), and infected with the resistant strain (R, red). The right panel illustrates patient movement between connected wards, regulated by the mobility parameter p_d.

If mobility is absent ($p_d = 0$), the system reduces to N independent copies of the single-ward model (10.3), and the stability and equilibrium results discussed in Section 10.1.1 apply node-wise. In the following, we investigate how patient mobility modifies the threshold behaviour of the system, using both the basic reproduction number R_0 and the critical transmission parameters.

10.3 Estimation of the basic reproduction number R_0

As discussed in Chapter 9, the basic reproduction number R_0 provides a threshold quantity that determines whether an infection can invade and persist. For simple models, R_0 can often be expressed as the ratio between a transmission parameter and a recovery parameter. For more complex systems, particularly in structured populations and multipathogen settings,

it is convenient to use the NGM formalism [28, 2, 75], adapted to discrete-time dynamics.

We first compute a *local* basic reproduction number R_{0i} for each ward considered in isolation (no mobility), and then derive the *global* reproduction number R_0 for the full metapopulation system with mobility.

10.3.1 Local R_0

Setting $p_d = 0$ decouples the wards, and the dynamics within ward i is governed by equations of the form (10.12), with a constant population n_i. Linearizing around the disease-free equilibrium and separating infection terms (matrix T_i) from transition terms (matrix Σ_i), we obtain for each ward i:

$$T_i = \begin{pmatrix} \beta_1 n_i & 0 \\ 0 & \beta_2 n_i \end{pmatrix}, \qquad \Sigma_i = \begin{pmatrix} 1-(\gamma+\tau_1+\tau_2+\sigma) & 0 \\ \sigma & 1-(\gamma+\tau_2) \end{pmatrix}. \tag{10.17}$$

The NGM for ward i is then:

$$K_i = T_i(\mathcal{I}-\Sigma_i)^{-1} = \begin{pmatrix} \dfrac{\beta_1 n_i}{\gamma+\tau_1+\tau_2+\sigma} & 0 \\ \dfrac{\beta_2 n_i \sigma}{(\gamma+\tau_1+\tau_2+\sigma)(\gamma+\tau_2)} & \dfrac{\beta_2 n_i}{\gamma+\tau_2} \end{pmatrix}. \tag{10.18}$$

The local basic reproduction number for ward i is given by the spectral radius of K_i:

$$R_{0i} = \max(\lambda_1, \lambda_2), \tag{10.19}$$

where:

$$\lambda_1 = \frac{\beta_1 n_i}{\gamma+\tau_1+\tau_2+\sigma}, \qquad \lambda_2 = \frac{\beta_2 n_i}{\gamma+\tau_2}. \tag{10.20}$$

The two eigenvalues correspond to the contributions of the susceptible and resistant strains, respectively. Even though the two strains are coupled via σ, the dominant eigenvalue still admits a clear interpretation in terms of effective reproduction numbers.

10.3.2 Global R_0

We now consider the full metapopulation system with $p_d > 0$. In this case, the infected subsystem is described by the vector:

$$\mathbf{X}(t) = \left(\rho_1^S(t), \ldots, \rho_N^S(t), \rho_1^R(t), \ldots, \rho_N^R(t)\right)^T \in \mathbb{R}^{2N}. \tag{10.21}$$

Linearizing equations (10.12) around the disease-free equilibrium ($\rho_i^{S*} \ll 1$, $\rho_i^{R*} \ll 1$) and keeping only linear terms in the infected fractions, we write the system in the standard NGM form:

$$\mathbf{X}(t+1) = (T + \Sigma)\,\mathbf{X}(t), \tag{10.22}$$

where T collects new infection terms and Σ contains transition terms (recovery and progression).

As shown in Chapter 9, the NGM is given by Eq. 9.10,and the basic reproduction number is defined as in Eq. 9.11, where $\rho(\cdot)$ denotes the spectral radius.

After linearization of the infection probabilities in equations (10.15)–(10.16) and some algebraic manipulation (details in Appendix A.3), the global transition matrix Σ takes the block form:

$$\Sigma = \begin{bmatrix} \mathcal{I}_N\left(1 - (\gamma + \tau_1 + \tau_2 + \sigma)\right) & \mathcal{O}_N \\ \mathcal{I}_N(\sigma) & \mathcal{I}_N\left(1 - (\gamma + \tau_2)\right) \end{bmatrix}_{2N \times 2N}, \tag{10.23}$$

while the global transmission matrix T has the structure:

$$T = \begin{bmatrix} T_S & \mathcal{O}_N \\ \mathcal{O}_N & T_R \end{bmatrix}_{2N \times 2N}, \tag{10.24}$$

where the (i,k)-th elements of T_S and T_R are:

$$(T_S)_{ik} = \beta_1\Big[(1-p_d)n_{k\to i} + p_d R_{ik} n_k^{\text{eff}}\Big], \qquad (T_R)_{ik} = \beta_2\Big[(1-p_d)n_{k\to i} + p_d R_{ik} n_k^{\text{eff}}\Big]. \tag{10.25}$$

Using (9.10) and the block structure of T and Σ, the NGM K can be written as:

$$K = \begin{bmatrix} K_{11} & \mathcal{O}_N \\ K_{21} & K_{22} \end{bmatrix}, \tag{10.26}$$

with diagonal blocks given by:

$$K_{11} = \frac{\beta_1}{\gamma + \tau_1 + \tau_2 + \sigma} M, \qquad K_{22} = \frac{\beta_2}{\gamma + \tau_2} M, \tag{10.27}$$

where the matrix $M = (M_{ij})$ encodes the mobility-weighted connectivity:

$$M_{ij} = (1 - p_d) n_{j \to i} + p_d R_{ij} n_j^{\text{eff}}. \tag{10.28}$$

The off-diagonal block K_{21} contains terms proportional to σ and does not affect the eigenvalues of K, because K is block lower triangular.

Since K is block lower triangular, its eigenvalues are the union of those of K_{11} and K_{22} [89]. It follows that:

$$R_0 = \rho(K) = \lambda_{\max}(M) \max \left\{ \frac{\beta_1}{\gamma + \tau_1 + \tau_2 + \sigma}, \frac{\beta_2}{\gamma + \tau_2} \right\}, \tag{10.29}$$

where $\lambda_{\max}(M)$ is the largest eigenvalue of M. Using $\beta_2 = (1 - c)\beta_1$, equation (10.29) can be rewritten as:

$$R_0 = \lambda_{\max}(M) \max \left\{ \frac{1}{\gamma + \tau_1 + \tau_2 + \sigma}, \frac{1 - c}{\gamma + \tau_2} \right\} \beta_1. \tag{10.30}$$

This expression highlights the interplay between hospital topology (via $\lambda_{\max}(M)$), epidemiological parameters, and resistance fitness cost. In particular, for fixed recovery and mobility parameters, R_0 grows linearly with β_1 and is modulated by both the structure of the hospital network and the relative fitness of the resistant strain.

10.4 Critical value of the transmission parameter

An alternative and complementary way to characterize the threshold behaviour of the system is to analyze the critical values of the transmission parameters β_1 and β_2 that separate disease-free from endemic regimes [41]. This approach is closely related to the NGM-based computation of R_0, and we show below that the two methods are consistent.

Near the epidemic threshold, the stationary fractions of infected individuals are small: $\rho_i^{S*} \ll 1$, $\rho_i^{R*} \ll 1$ for all i. Linearizing the infection probabilities in (10.15) and using the definition of $n_{j \to i}$ in (10.16), we can express the infection terms in the form (A.2):

$$\Pi_i^{X \to S}(t) \approx \beta_1 (M \vec{\rho}^{S*})_i, \qquad \Pi_i^{X \to R}(t) \approx \beta_2 (M \vec{\rho}^{R*})_i, \tag{10.31}$$

where $\vec{\rho}^{S*}, \vec{\rho}^{R*} \in \mathbb{R}^N$ are the vectors of stationary infected fractions, and M is the same mobility-weighted matrix as in (10.28). Substituting (10.31) into the stationary version of the linearized system (10.12), we obtain:

$$\begin{aligned} (\gamma + \tau_1 + \tau_2 + \sigma)\, \rho_i^{S*} &\approx \beta_1 (M\vec{\rho}^{S*})_i, \\ (\gamma + \tau_2)\, \rho_i^{R*} - \sigma \rho_i^{S*} &\approx \beta_2 (M\vec{\rho}^{R*})_i. \end{aligned} \tag{10.32}$$

To identify the critical transmission values, we consider small perturbations around the disease-free equilibrium of the form:

$$\vec{\rho}^{S*} \approx \varepsilon \vec{v}_S, \qquad \vec{\rho}^{R*} \approx \varepsilon \vec{v}_R, \tag{10.33}$$

where $\varepsilon \ll 1$ and $\vec{v}_S$, $\vec{v}_R$ are eigenvectors of M associated with an eigenvalue λ. Neglecting the term proportional to $\sigma \vec{v}_S$ in the second equation (consistent with σ small), we obtain the scalar conditions:

$$(\gamma + \tau_1 + \tau_2 + \sigma) \approx \beta_1 \lambda, \qquad (\gamma + \tau_2) \approx \beta_2 \lambda. \tag{10.34}$$

At the threshold, the relevant eigenvalue is $\lambda_{\max}(M)$, and the critical transmission parameters are therefore:

$$\beta_{1c} = \frac{\gamma + \tau_1 + \tau_2 + \sigma}{\lambda_{\max}(M)}, \qquad \beta_{2c} = \frac{\gamma + \tau_2}{\lambda_{\max}(M)}. \tag{10.35}$$

Recalling that $\beta_2 = (1 - c)\beta_1$, we can rewrite β_{2c} as:

$$\beta_{1c} = \frac{\gamma + \tau_2}{(1 - c)\lambda_{\max}(M)}. \tag{10.36}$$

The critical value of β_1 is then given by:

$$\beta_{1c} = \min \left\{ \frac{\gamma + \tau_1 + \tau_2 + \sigma}{\lambda_{\max}(M)},\ \frac{\gamma + \tau_2}{(1 - c)\lambda_{\max}(M)} \right\}. \tag{10.37}$$

In numerical simulations, the relevant threshold is almost always the second expression in (10.37), which is associated with the resistant strain and the fitness cost c.

Biologically, equations (10.35)–(10.37) show that the epidemic threshold arises from the interaction between recovery and treatment rates at the ward level, the fitness cost associated with resistance, and the structural properties of the hospital network. In particular, the largest eigenvalue

of the mobility-weighted matrix M quantifies how efficiently patient movement couples the wards, thereby amplifying or attenuating the potential for pathogen transmission. When mobility channels align with high local transmission potential, the resulting increase in $\lambda_{\max}(M)$ decreases the critical value of β_1, making it easier for both sensitive and resistant strains to invade. Conversely, strong recovery or treatment rates, together with a substantial fitness cost of resistance, shift the system toward higher critical thresholds, limiting the ability of resistant strains to persist within the hospital network.

10.5 Connection between the two methods

We now make explicit the connection between the NGM-based computation of R_0 and the critical-transmission approach.

Starting from the expression of the NGM K in equation (10.26) and using the block forms of K_{11} and K_{22} in (10.27), we obtained:

$$R_0 = \rho(K) = \lambda_{\max}(M) \max\left\{\frac{\beta_1}{\gamma + \tau_1 + \tau_2 + \sigma}, \frac{\beta_2}{\gamma + \tau_2}\right\}. \tag{10.38}$$

Setting $R_0 = 1$ and solving for β_1 (A.4) yields:

$$\beta_{1c} = \frac{1}{\lambda_{\max}(M)} \min\left\{\gamma + \tau_1 + \tau_2 + \sigma, \frac{\gamma + \tau_2}{1 - c}\right\}, \tag{10.39}$$

which is exactly the expression in (10.37).

Thus, the condition $R_0 = 1$ and the condition $\beta_1 = \beta_{1c}$ are mathematically equivalent.

The two approaches, one formulated in terms of an eigenvalue problem for the NGM, the other in terms of critical transmission parameters derived from the steady-state linearization, provide consistent and complementary descriptions of the epidemic threshold.

From a modeling perspective, this equivalence shows that the epidemic threshold is primarily shaped by the largest eigenvalue of the matrix M, which encodes the effective connectivity of the hospital network and quantifies how strongly mobility patterns facilitate the spread of infection. As a consequence, the system's threshold behaviour can be modulated in two conceptually distinct ways: by acting on local epidemiological parameters, such as recovery probabilities, treatment policies, or the fitness cost of resistance,

or by modifying mobility structures, for instance through adjustments in patient transfer protocols. These two levers, acting respectively at the ward and network scales, jointly determine whether resistant or sensitive strains can successfully invade and persist.

10.6 Final remarks

The XSR model provides a mathematically consistent framework for describing the competition between antibiotic-sensitive and antibiotic-resistant strains in a spatially structured hospital environment. By combining a three-compartment structure (X, S, R) with a metapopulation representation of patient mobility, the model captures mechanisms that are essential in real hospital settings: ward-specific contact patterns, strain-dependent transmissibility, selective pressures induced by antibiotic use, and mobility constraints imposed by infection control policies.

The local stability analysis of the single-ward system clarifies the conditions under which the susceptible or the resistant strain dominates and identifies the narrow parameter regime in which coexistence is possible. Extending the model to a metapopulation system requires incorporating patient mobility; this is achieved through the derivation of the basic reproduction number R_0 via the NGM and through the identification of critical transmission thresholds. Both approaches converge to the same condition, showing that the epidemic threshold is governed by the largest eigenvalue of the mobility-weighted connectivity matrix M.

The model introduced in this chapter is sufficiently general to be adapted to different mobility patterns, antibiotic policies, and pathogen-specific dynamics. At the same time, it exposes structural limitations that must be kept in mind: it assumes homogeneous mixing within wards, deterministic dynamics, and does not explicitly model true co-colonization or stochastic extinction events. These limitations motivate the more refined, data-driven analyzes developed in the following chapters, where the XSR framework will serve as a reference model for interpreting simulation results and guiding the design of control strategies.

Chapter 11

A metapopulation model with two competitive strains

In this chapter, we present a discrete-time compartmental metapopulation model that simultaneously tracks the spread of two infectious agents across a network of interconnected subpopulations. The model accounts for both single infections and coinfections, representing individuals in one of four compartments: susceptible to both agents (SS), infected only by the first agent (IS), infected only by the second agent (SI), or coinfected by both agents (II). Pathogen interactions are incorporated through cost parameters that modulate transmission and coinfection probabilities, while mobility restrictions are introduced in a compartment-specific way to mimic isolation protocols or reduced mobility due to disease severity.

This model allows us to investigate how strain competition, penalties for coinfection, and spatial mobility restrictions shape the long-term prevalence of each infection class. Through numerical simulations and spectral analyses, we explore the sensitivity of the dynamics to key epidemiological parameters, including infection rates, recovery rates, interaction cost factors, and the effectiveness of isolation policies. Although the model is general and can be applied to different structured settings (such as cities or hospital networks), we focus on an interpretation in the context of nosocomial transmission and AMR.

11.1 Model

The metapopulation structure underlying the model is based on a real mobility network, in which each node represents a subpopulation and edges encode the probability of movement between nodes. In particular, we consider a network derived from recurrent mobility patterns in an urban setting (specifically, commuting flows between districts of the city of Cali), as in reaction–diffusion studies of epidemic spreading on networks [41]. The resulting weighted adjacency matrix $W = (W_{ij})$ describes the baseline connectivity between nodes, while a row-stochastic matrix

$$R_{ij} = \frac{W_{ij}}{\sum_k W_{ik}} \tag{11.1}$$

is used to model the probabilistic movement of individuals.

In the hospital interpretation adopted later in this chapter, each node will represent a ward, and R_{ij} will correspond to normalized patient transfer probabilities between wards. The same network structure is used both in the full SIIS model with coinfection and in the simplified two-strain model employed for the analytical derivation of the threshold behaviour.

To describe the model we consider a population divided into N subpopulations (nodes), each of which is internally well-mixed. Two infectious agents, denoted A_1 and A_2, can circulate simultaneously. Each individual belongs to exactly one of four compartments:

- SS: susceptible to both agents A_1 and A_2;
- IS: infected only by agent A_1;
- SI: infected only by agent A_2;
- II: coinfected by both A_1 and A_2.

Transmission occurs through direct or indirect contact with infected individuals in the same node and is governed by a baseline infection probability λ and interaction cost factors q and ν: $q \in [0, 1]$ modulates the transmissibility of the second agent relative to the first (e.g. resistant vs. sensitive strain), while ν is a parameter that regulate the possibility to have cooperative disease (if $\nu > 1$) or competitive diseases ($\nu \in [0, 1]$) [104]. In our case, we consider the second scenario: ν modulates the probability of acquiring a

second infection given an existing single infection, i.e. a coinfection penalty. Infected individuals may recover with probability μ per time step. Recovery is modeled as a stepwise reduction of infectious load: individuals in compartment II can first move to IS or SI, and then to SS; no direct transitions between IS and SI or between II and SS are assumed.

The model admits a natural interpretation in terms of nosocomial transmission of bacteria, distinguishing between strains sensitive (S) and resistant (R) to a given antimicrobial agent. In this framework, the compartment SS corresponds to patients who are neither colonised nor infected, while IS denotes individuals infected with bacteria that remain sensitive to antibiotic treatment. Conversely, the SI compartment represents patients colonised or infected by a resistant strain, reflecting reduced efficacy of the antimicrobial therapy. Finally, the II compartment accounts for mixed colonisation or coinfection, in which both sensitive and resistant strains coexist within the same host.

Healthy patients in SS can acquire a sensitive infection with probability λ, or a resistant infection with probability $q\lambda$, where $0 < q < 1$ reflects the lower transmissibility typically associated with resistant strains. Patients already infected with one strain (IS or SI) can acquire the other, moving to II, with probability reduced by the factor ν.

Spontaneous or treatment-induced recovery is captured by a probability μ, which governs partial or complete clearance of the infection. The absence of direct transitions between IS and SI or between II and SS is consistent with the assumption that recovery and loss of colonisation occur progressively.

The model is schematically represented in Fig. 11.1.

Starting from this model, we now move to a metapopulation structure. As we mentioned before, the population is distributed over N nodes, which may represent hospital wards or geographic areas. Individuals can move between nodes according to a mobility parameter p and compartment-specific reduction factors.

Let $\rho_i^{SS}(t)$, $\rho_i^{IS}(t)$, $\rho_i^{SI}(t)$ and $\rho_i^{II}(t)$ denote the fractions of individuals in node i at time t belonging to the four compartments, with

$$\rho_i^{SS}(t) + \rho_i^{IS}(t) + \rho_i^{SI}(t) + \rho_i^{II}(t) = 1. \tag{11.2}$$

Mobility is implemented as follows:

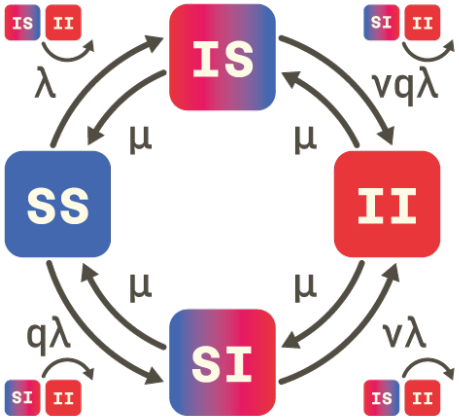


Figure 11.1: **SIIS model:** patients can be healthy (SS), infected by a susceptible strain (IS), infected by a resistant strain (SI) or infected by both strains (II).

- SS individuals move with baseline probability p;
- IS individuals move with reduced probability αp;
- SI individuals move with reduced probability βp;
- II individuals, being coinfected, are subject to cumulative restrictions and move with probability $\alpha\beta p$.

Here $\alpha, \beta \in [0, 1]$ quantify the strength of mobility reduction due to infection or isolation policies. For instance, $\alpha < 1$ may represent moderate containment measures for patients carrying the sensitive strain, while $\beta < 1$ may represent more stringent isolation protocols for resistant carriers.

11.1.1 Model equations

We now write the discrete-time equations governing the evolution of the infected compartments. For node i, the fractions of individuals in IS, SI,

and II at time $t+1$ are given by:

$$\begin{aligned}
\rho_i^{IS}(t+1) &= \rho_i^{IS}(t)(1-\mu)\big[1-\Pi_i^{IS\to II}(t)\big] + \rho_i^{II}(t)(1-\mu)\mu + \\
&\quad + \rho_i^{SS}(t)\,\Pi_i^{SS\to IS}(t)\big[1-\Pi_i^{SS\to SI}(t)\big], \\
\rho_i^{SI}(t+1) &= \rho_i^{SI}(t)(1-\mu)\big[1-\Pi_i^{SI\to II}(t)\big] + \rho_i^{II}(t)(1-\mu)\mu + \\
&\quad + \rho_i^{SS}(t)\,\Pi_i^{SS\to SI}(t)\big[1-\Pi_i^{SS\to IS}(t)\big], \\
\rho_i^{II}(t+1) &= \rho_i^{II}(t)(1-\mu)^2 + \rho_i^{SI}(t)(1-\mu)\,\Pi_i^{SI\to II}(t) + \\
&\quad + \rho_i^{IS}(t)(1-\mu)\,\Pi_i^{IS\to II}(t),
\end{aligned} \tag{11.3}$$

where $\Pi_i^{A\to B}(t)$ denotes the probability that an individual associated with node i and currently in compartment A moves to compartment B during the time step $[t, t+1]$. The infection-related transition probabilities for residents of node i at time t are:

$$\begin{aligned}
\Pi_i^{SS\to IS}(t) &= (1-p)\,P_i^{SS\to IS}(t) + p\sum_{j=1}^{N} R_{ij} P_j^{SS\to IS}(t), \\
\Pi_i^{SS\to SI}(t) &= (1-p)\,P_i^{SS\to SI}(t) + p\sum_{j=1}^{N} R_{ij} P_j^{SS\to SI}(t), \\
\Pi_i^{IS\to II}(t) &= (1-\alpha p)\,P_i^{IS\to II}(t) + \alpha p\sum_{j=1}^{N} R_{ij} P_j^{IS\to II}(t), \\
\Pi_i^{SI\to II}(t) &= (1-\beta p)\,P_i^{SI\to II}(t) + \beta p\sum_{j=1}^{N} R_{ij} P_j^{SI\to II}(t).
\end{aligned} \tag{11.4}$$

Here, $P_i^{A\to B}(t)$ denotes the probability that a susceptible individual placed in node i at time t acquires infection A_1 or A_2 (or the second agent in the case of coinfection), conditional on remaining in that node. These probabilities depend on the number of infected individuals present in node

i after mobility. Then:

$$
\begin{aligned}
P_i^{SS\to IS}(t) &= \left[1-\prod_{j=1}^{N}(1-\lambda)^{n_{j\to i}^{IS}(t)+n_{j\to i}^{II}(t)}(1-q\lambda)^{n_{j\to i}^{SI}(t)+n_{j\to i}^{II}(t)}\right] f_{SS\to IS}(t),\\
P_i^{SS\to SI}(t) &= \left[1-\prod_{j=1}^{N}(1-\lambda)^{n_{j\to i}^{IS}(t)+n_{j\to i}^{II}(t)}(1-q\lambda)^{n_{j\to i}^{SI}(t)+n_{j\to i}^{II}(t)}\right] f_{SS\to SI}(t),\\
P_i^{SI\to II}(t) &= 1-\prod_{j=1}^{N}(1-\nu\lambda)^{n_{j\to i}^{IS}(t)+n_{j\to i}^{II}(t)},\\
P_i^{IS\to II}(t) &= 1-\prod_{j=1}^{N}(1-\nu q\lambda)^{n_{j\to i}^{SI}(t)+n_{j\to i}^{II}(t)}.
\end{aligned}
\tag{11.5}
$$

The functions $f_{SS\to IS}(t)$ and $f_{SS\to SI}(t)$ represent the probability that a susceptible individual in node i first encounters an infective carrying agent A_1 or A_2, respectively:

$$
\begin{aligned}
f_{SS\to IS}(t) &= \frac{n_i^{IS,\mathrm{eff}}(t)+n_i^{II,\mathrm{eff}}(t)}{n_i^{IS,\mathrm{eff}}(t)+n_i^{II,\mathrm{eff}}(t)+q\big(n_i^{SI,\mathrm{eff}}(t)+n_i^{II,\mathrm{eff}}(t)\big)},\\
f_{SS\to SI}(t) &= \frac{q\big(n_i^{SI,\mathrm{eff}}(t)+n_i^{II,\mathrm{eff}}(t)\big)}{n_i^{IS,\mathrm{eff}}(t)+n_i^{II,\mathrm{eff}}(t)+q\big(n_i^{SI,\mathrm{eff}}(t)+n_i^{II,\mathrm{eff}}(t)\big)}.
\end{aligned}
\tag{11.6}
$$

They depend on the effective number of infected individuals present in node i. In particular, the number of individuals in compartment $A \in \{IS, SI, II\}$ moving from node j to node i at time t is denoted by $n_{j\to i}^{A}(t)$ and is given by:

$$
\begin{aligned}
n_{j\to i}^{IS}(t) &= \delta_{ij}\big[(1-\alpha p)\rho_j^{IS}(t)n_j\big] + R_{ji}\,\alpha p\,\rho_j^{IS}(t)n_j,\\
n_{j\to i}^{SI}(t) &= \delta_{ij}\big[(1-\beta p)\rho_j^{SI}(t)n_j\big] + R_{ji}\,\beta p\,\rho_j^{SI}(t)n_j,\\
n_{j\to i}^{II}(t) &= \delta_{ij}\big[(1-\alpha\beta p)\rho_j^{II}(t)n_j\big] + R_{ji}\,\alpha\beta p\,\rho_j^{II}(t)n_j,
\end{aligned}
\tag{11.7}
$$

where n_j is the population size of node j and δ_{ij} is the Kronecker delta. The effective numbers of infected individuals present in node i at time t are

then:

$$
\begin{aligned}
n_i^{IS,\text{eff}}(t) &= (1-\alpha p)\rho_i^{IS}(t)n_i + \alpha p \sum_{j=1}^{N} R_{ji}\rho_j^{IS}(t)n_j, \\
n_i^{SI,\text{eff}}(t) &= (1-\beta p)\rho_i^{SI}(t)n_i + \beta p \sum_{j=1}^{N} R_{ji}\rho_j^{SI}(t)n_j, \\
n_i^{II,\text{eff}}(t) &= (1-\alpha\beta p)\rho_i^{II}(t)n_i + \alpha\beta p \sum_{j=1}^{N} R_{ji}\rho_j^{II}(t)n_j.
\end{aligned}
\tag{11.8}
$$

where α and β are the mobility reduction parameters.

Equations (11.3)–(11.8) define the full SIIS metapopulation model, including coinfection. In the next section, we focus on a simplified version of the model (no coinfection) to derive explicit expressions for the epidemic threshold and R_0.

11.2 Estimation of R_0 and critical transmission values in the case $\nu = 0$

As explained in Chapter 9, the basic reproduction number R_0 provides a threshold criterion for the persistence of infection. For complex structured systems, R_0 can be computed using the NGM method by separating new infection terms from transition terms in the linearised dynamics around the disease-free equilibrium.

To obtain tractable expressions, we consider a simplified version of the SIIS model in which coinfection is not allowed, i.e. we set $\nu = 0$ and $\rho_i^{II}(t) \equiv 0$. The system then reduces to a two-strain metapopulation model with compartments SS, IS, and SI.

In the absence of coinfection ($\nu = 0$), the dynamics for IS and SI in node i are:

$$
\begin{aligned}
\rho_i^{IS}(t+1) &= \rho_i^{IS}(t)(1-\mu) + \rho_i^{SS}(t)\,\Pi_i^{SS\to IS}(t)\big[1-\Pi_i^{SS\to SI}(t)\big], \\
\rho_i^{SI}(t+1) &= \rho_i^{SI}(t)(1-\mu) + \rho_i^{SS}(t)\,\Pi_i^{SS\to SI}(t)\big[1-\Pi_i^{SS\to IS}(t)\big],
\end{aligned}
\tag{11.9}
$$

where the metapopulation infection probabilities are:

$$\begin{aligned}\Pi_i^{SS\to IS}(t) &= (1-p)\,P_i^{SS\to IS}(t) + p\sum_{j=1}^{N} R_{ij}P_j^{SS\to IS}(t),\\ \Pi_i^{SS\to SI}(t) &= (1-p)\,P_i^{SS\to SI}(t) + p\sum_{j=1}^{N} R_{ij}P_j^{SS\to SI}(t).\end{aligned} \tag{11.10}$$

The local infection probabilities can be written in terms of the effective numbers $n_i^{IS,\text{eff}}$ and $n_i^{SI,\text{eff}}$ defined in (11.8). Assuming z potentially infectious contacts per individual per time step, we have:

$$\begin{aligned}P_i^{SS\to IS}(t) &= \left[1-(1-\lambda)^{zn_i^{IS,\text{eff}}(t)}(1-q\lambda)^{zn_i^{SI,\text{eff}}(t)}\right] f_{SS\to IS}(t),\\ P_i^{SS\to SI}(t) &= \left[1-(1-\lambda)^{zn_i^{IS,\text{eff}}(t)}(1-q\lambda)^{zn_i^{SI,\text{eff}}(t)}\right] f_{SS\to SI}(t),\end{aligned} \tag{11.11}$$

where, with $\rho_i^{II}\equiv 0$,

$$\begin{aligned}f_{SS\to IS}(t) &= \frac{n_i^{IS,\text{eff}}(t)}{n_i^{IS,\text{eff}}(t)+qn_i^{SI,\text{eff}}(t)},\\ f_{SS\to SI}(t) &= \frac{qn_i^{SI,\text{eff}}(t)}{n_i^{IS,\text{eff}}(t)+qn_i^{SI,\text{eff}}(t)}.\end{aligned} \tag{11.12}$$

The effective numbers in (11.8) reduce to:

$$\begin{aligned}n_i^{IS,\text{eff}}(t) &= (1-\alpha p)n_i\rho_i^{IS}(t) + \alpha p\sum_{j=1}^{N} R_{ji}n_j\rho_j^{IS}(t),\\ n_i^{SI,\text{eff}}(t) &= (1-\beta p)n_i\rho_i^{SI}(t) + \beta p\sum_{j=1}^{N} R_{ji}n_j\rho_j^{SI}(t).\end{aligned} \tag{11.13}$$

We now linearise the simplified system around the disease-free equilibrium, where $\rho_i^{IS},\rho_i^{SI}\ll 1$ and $\rho_i^{SS}\approx 1$. Let ρ_i^{*IS} and ρ_i^{*SI} denote the small perturbations around the disease-free equilibrium in IS and SI, respectively.

Using the approximation $(1-a)^x\approx x\log(1-a)$ for small x, we obtain from (11.11):

$$\begin{aligned}P_i^{SS\to IS}(t) &\approx -z\log(1-\lambda)\,n_i^{IS,\text{eff}}(t),\\ P_i^{SS\to SI}(t) &\approx -z\log(1-q\lambda)\,n_i^{SI,\text{eff}}(t),\end{aligned} \tag{11.14}$$

Substituting (11.14) into (11.10), the linearised metapopulation infection probabilities become:

$$\begin{aligned}\Pi_i^{*SS\to IS} &\approx -z\log(1-\lambda)\left[(1-p)n_i^{IS,\text{eff}} + p\sum_{j=1}^{N} R_{ij}n_j^{IS,\text{eff}}\right],\\ \Pi_i^{*SS\to SI} &\approx -z\log(1-q\lambda)\left[(1-p)n_i^{SI,\text{eff}} + p\sum_{j=1}^{N} R_{ij}n_j^{SI,\text{eff}}\right].\end{aligned} \tag{11.15}$$

Since $\rho_i^{SS} \approx 1$ and terms of order $\Pi_i^{SS\to IS}\Pi_i^{SS\to SI}$ are negligible, the linearised version of (11.9) for the perturbations reads:

$$\begin{aligned}\rho_i^{*IS} &\approx \rho_i^{*IS}(1-\mu) + \Pi_i^{*SS\to IS},\\ \rho_i^{*SI} &\approx \rho_i^{*SI}(1-\mu) + \Pi_i^{*SS\to SI}.\end{aligned} \tag{11.16}$$

To write the system in matrix form, we define the diagonal matrix $\mathbf{N} = \text{diag}(n_1,\ldots,n_N)$ and the vectors $\boldsymbol{\rho}^{IS} = (\rho_1^{IS},\ldots,\rho_N^{IS})^T$, $\boldsymbol{\rho}^{SI} = (\rho_1^{SI},\ldots,\rho_N^{SI})^T$. Using (11.13), we can express the effective populations as:

$$\begin{aligned}\mathbf{n}^{IS,\text{eff}} &= \left[(1-\alpha p)\mathbf{N} + \alpha p R^T\mathbf{N}\right]\boldsymbol{\rho}^{IS} \equiv A^{(\alpha)}(p)\,\boldsymbol{\rho}^{IS},\\ \mathbf{n}^{SI,\text{eff}} &= \left[(1-\beta p)\mathbf{N} + \beta p R^T\mathbf{N}\right]\boldsymbol{\rho}^{SI} \equiv A^{(\beta)}(p)\,\boldsymbol{\rho}^{SI}.\end{aligned} \tag{11.17}$$

Similarly, we define the mobility mixing matrix:

$$M(p) = (1-p)\mathcal{I}_N + pR. \tag{11.18}$$

Using (11.14) and (11.18), the vector of infection probabilities can be written as:

$$\begin{aligned}\mathbf{P}^{SS\to IS} &\approx -z\log(1-\lambda)\,\mathbf{n}^{IS,\text{eff}},\\ \mathbf{P}^{SS\to SI} &\approx -z\log(1-q\lambda)\,\mathbf{n}^{SI,\text{eff}},\end{aligned} \tag{11.19}$$

and the corresponding metapopulation transition probabilities as:

$$\begin{aligned}\boldsymbol{\Pi}^{SS\to IS} &\approx M(p)\,\mathbf{P}^{SS\to IS} = -z\log(1-\lambda)\,M(p)A^{(\alpha)}(p)\,\boldsymbol{\rho}^{IS},\\ \boldsymbol{\Pi}^{SS\to SI} &\approx M(p)\,\mathbf{P}^{SS\to SI} = -z\log(1-q\lambda)\,M(p)A^{(\beta)}(p)\,\boldsymbol{\rho}^{SI}.\end{aligned} \tag{11.20}$$

Substituting into the linearised dynamics (11.16) and writing in vector form, we obtain:

$$\begin{aligned}\boldsymbol{\rho}^{*IS} &\approx (1-\mu)\,\boldsymbol{\rho}^{*IS} - z\log(1-\lambda)\,M(p)A^{(\alpha)}(p)\,\boldsymbol{\rho}^{*IS},\\ \boldsymbol{\rho}^{*SI} &\approx (1-\mu)\,\boldsymbol{\rho}^{*SI} - z\log(1-q\lambda)\,M(p)A^{(\beta)}(p)\,\boldsymbol{\rho}^{*SI}.\end{aligned} \tag{11.21}$$

Comparing (11.21) with the general NGM form in discrete time

$$\mathbf{X}(t+1) = (T+\Sigma)\mathbf{X}(t),$$

we derive, for each strain,

$$\begin{aligned} \Sigma^{IS} &= (1-\mu)\,\mathcal{I}_N, \qquad & T^{IS} &= -z\log(1-\lambda)\,M(p)A^{(\alpha)}(p), \\ \Sigma^{SI} &= (1-\mu)\,\mathcal{I}_N, \qquad & T^{SI} &= -z\log(1-q\lambda)\,M(p)A^{(\beta)}(p). \end{aligned} \tag{11.22}$$

The NGM blocks for the two strains are then:

$$\begin{aligned} K^{IS} &= T^{IS}(\mathcal{I}_N - \Sigma^{IS})^{-1} = \frac{-z\log(1-\lambda)}{\mu} M(p)A^{(\alpha)}(p), \\ K^{SI} &= T^{SI}(\mathcal{I}_N - \Sigma^{SI})^{-1} = \frac{-z\log(1-q\lambda)}{\mu} M(p)A^{(\beta)}(p). \end{aligned} \tag{11.23}$$

The full NGM is block diagonal in this simplified setting, with K^{IS} and K^{SI} on the diagonal. The basic reproduction number is then given by the spectral radius of the full matrix:

$$\begin{aligned} R_0 &= \rho(K) = \max\left\{\rho(K^{IS}),\, \rho(K^{SI})\right\} = \\ &= \max\left\{\tfrac{-z\log(1-\lambda)}{\mu}\,\rho\big(M(p)A^{(\alpha)}(p)\big),\, \tfrac{-z\log(1-q\lambda)}{\mu}\,\rho\big(M(p)A^{(\beta)}(p)\big)\right\}. \end{aligned} \tag{11.24}$$

The epidemic threshold corresponds to the condition $R_0 = 1$. From equation (11.24), the critical transmission probabilities λ_c^α and λ_c^β associated with the two strains satisfy:

$$\begin{aligned} \rho\big(M(p)A^{(\alpha)}(p)\big) &= -\frac{\mu}{z\log(1-\lambda_c^\alpha)}, \\ \rho\big(M(p)A^{(\beta)}(p)\big) &= -\frac{\mu}{z\log(1-q\lambda_c^\beta)}. \end{aligned} \tag{11.25}$$

Solving for λ_c^α and λ_c^β yields:

$$\begin{aligned} \lambda_c^\alpha &= 1 - \exp\left(-\frac{\mu}{z\,\rho\big(M(p)A^{(\alpha)}(p)\big)}\right), \\ \lambda_c^\beta &= \frac{1}{q}\left[1 - \exp\left(-\frac{\mu}{z\,\rho\big(M(p)A^{(\beta)}(p)\big)}\right)\right]. \end{aligned} \tag{11.26}$$

The global critical value of λ is then given by:

$$\lambda_c = \min\{\lambda_c^\alpha,\, \lambda_c^\beta\}, \tag{11.27}$$

which identifies the smallest transmission probability at which at least one of the two strains is able to invade the metapopulation system.

The expression (11.24) makes explicit how the epidemic threshold depends jointly on local epidemiological parameters (λ, q, μ), the mobility level p, and the network structure encoded in $M(p)$ and the mobility-restricted matrices $A^{(\alpha)}(p)$ and $A^{(\beta)}(p)$.

Finally, if we want to find the minimum value of the mobility p^* that minimizes R_0, fixed λ, we have to calculate:

$$\left.\frac{\partial(R_0)}{\partial p}\right|_{p^*} = 0 \tag{11.28}$$

11.3 Estimation of the invasion reproduction number $R_{\mathbf{inv}}$ in the case $\nu = 0$

As discussed in Chapter 9, NGM formalism can be extended beyond the basic reproduction number R_0 to describe invasion phenomena in multi-strain or multi-pathogen systems. While R_0 is defined by linearising the system around the disease-free equilibrium, R_{inv} is obtained by linearising the dynamics of the invading strain around an endemic equilibrium of the resident strain. In discrete time metapopulation models, this leads to Eq. (9.15).

In the simplified SIIS model without coinfection, we have two strains: the "sensitive" strain, associated with compartment IS, and the "resistant" strain, associated with compartment SI. Their NGMs at the disease-free equilibrium have been derived in Section 11.2 and are given by Eq. (11.23) where $M(p)$ encodes the baseline mobility pattern, while $A^{(\alpha)}(p)$ and $A^{(\beta)}(p)$ capture the effect of mobility reductions α and β on the two strains.

Let $\mathbf{S}^*_{IS}$ denote the vector of susceptible fractions at the endemic equilibrium where only the sensitive strain (IS) is present, and analogously $\mathbf{S}^*_{SI}$ the susceptible fractions at the endemic equilibrium where only the resistant strain (SI) is present. In both cases, the endemic equilibrium is characterized by the balance between new infections and recoveries for the resident strain, and satisfies the corresponding threshold condition:

$$\rho\left(D(\mathbf{S}^*_{IS})\,K^{IS}\right) = 1, \qquad \rho\left(D(\mathbf{S}^*_{SI})\,K^{SI}\right) = 1. \tag{11.29}$$

Given these equilibrium susceptible profiles, the invasion reproduction numbers for the two possible invasion scenarios are:

$$\begin{aligned} R_{(SI\,|\,IS)} &= \rho\,\left(D(\mathbf{S}^*_{IS})\,K^{SI}\right), \\ R_{(IS\,|\,SI)} &= \rho\,\left(D(\mathbf{S}^*_{SI})\,K^{IS}\right). \end{aligned} \tag{11.30}$$

Here $R_{(SI\,|\,IS)}$ measures the ability of the resistant strain (SI) to invade a metapopulation where the sensitive strain (IS) has already reached its endemic equilibrium, while $R_{(IS\,|\,SI)}$ measures the ability of the sensitive strain (IS) to invade a metapopulation where the resistant strain (SI) is endemic.

Biologically, $R_{(SI\,|\,IS)} > 1$ indicates that the resistant strain can establish itself in a hospital network where the sensitive strain is already circulating, despite its reduced transmissibility (controlled by q) and possibly stronger isolation (controlled by β). Conversely, $R_{(SI\,|\,IS)} < 1$ implies that the combination of fitness cost, recovery, and mobility restrictions is sufficient to prevent the spread of resistance on the background of an endemic sensitive strain. The symmetric interpretation holds for $R_{(IS\,|\,SI)}$.

From a modelling perspective, the pair

$$\left(R_{(SI\,|\,IS)},\, R_{(IS\,|\,SI)}\right)$$

provides a natural framework to classify the qualitative outcomes of competition between sensitive and resistant strains:

- if $R_{(SI\,|\,IS)} < 1$ and $R_{(IS\,|\,SI)} > 1$, only the sensitive strain can invade and persist;
- if $R_{(SI\,|\,IS)} > 1$ and $R_{(IS\,|\,SI)} < 1$, the resistant strain excludes the sensitive one;
- if both invasion reproduction numbers exceed one, mutual invasion is possible and coexistence scenarios may arise, depending on nonlinear feedbacks and initial conditions;
- if both are below one, neither strain can invade the endemic state of the other, and the system exhibits strong priority effects (the first established strain tends to persist).

In the following section, these invasion criteria will be evaluated numerically for the SIIS metapopulation model, highlighting how infection parameters (λ, q, μ), mobility restrictions (α, β, p) and network structure jointly shape the competitive balance between sensitive and resistant strains.

11.4 Derivation of R_0 and R_{inv} in the case $\nu \neq 0$

Allowing transitions to the co-infected compartment II does not modify the basic reproduction number R_0. The reason is that, at the disease-free equilibrium, co-infections satisfy $\rho^{II} = \mathcal{O}(\rho^{IS}\rho^{SI})$ and therefore $n_i^{II,\text{eff}} = \mathcal{O}(\varepsilon^2)$ for perturbations of size ε. All ν-dependent terms vanish in the linearisation of the system ($\mathcal{O}(\varepsilon^2)$), so the disease-free equilibrium Jacobian and its leading eigenvalue are identical to the case $\nu = 0$. Hence:

$$R_0(\lambda, p) = \max\left\{ \frac{\tau(\lambda)}{\mu}\, \rho\, (M(p)A_\alpha(p))\,,\, \frac{\tau_q(\lambda)}{\mu}\, \rho\, (M(p)A_\beta(p)) \right\}, \quad (11.31)$$

with $\tau(\lambda) = -z\ln(1-\lambda)$ and $\tau_q(\lambda) = -z\ln(1-q\lambda)$.

In contrast, ν does affect invasion. When one strain (say, IS) reaches an endemic equilibrium, co-infections generated by the resident enter the dynamics experienced by an invading strain (SI) introduced at very low prevalence, and modify its linearised growth rate. Let:

$$x(t) = \begin{bmatrix} \rho^{SI}(t) \\ \\ \delta\rho^{II}(t) \end{bmatrix} \quad (11.32)$$

collect the state variables associated with the invading pathogen. Its linearised dynamics around the resident equilibrium can be written as:

$$x(t+1) = (T_{\text{inv}}(\nu) + \Sigma_{\text{res}}(\nu))\, x(t), \quad (11.33)$$

where $\Sigma_{\text{res}}(\nu)$ denotes the transition matrix of the invading subsystem, evaluated at the resident equilibrium, and $T_{\text{inv}}(\nu)$ collects the corresponding new-infection terms.

The invasion next-generation operator is:

$$K_{\text{inv}}(\nu) = T_{\text{inv}}(\nu)\, (I - \Sigma_{\text{res}}(\nu))^{-1}. \quad (11.34)$$

In analogy with case with $\nu = 0$, the effect of the resident equilibrium enters the invasion dynamics exclusively through the availability of susceptible

individuals, which rescales the transmission terms but does not affect transitions among infected compartments. Accordingly, the linearised dynamics of the invading pathogen can be written as:

$$x(t+1) = \Big[\Sigma_{\text{res}}(\nu) + D\Big(\mathbf{S}^*_{(\text{res})}\Big)\, T_{\text{inv}}(\nu)\Big]\, x(t). \tag{11.35}$$

This implies that the effective invasion operator is:

$$K^{\text{eff}}_{\text{inv}}(\nu) = D\Big(\mathbf{S}^*_{(\text{res})}\Big)\; T_{\text{inv}}(\nu)\; (I - \Sigma_{\text{res}}(\nu))^{-1} = D\Big(\mathbf{S}^*_{(\text{res})}\Big)\; K_{\text{inv}}(\nu). \tag{11.36}$$

For an invader introduced into the endemic equilibrium of the resident strain, let $\mathbf{S}^*_{(\text{res})}$ denote the susceptible fractions at that equilibrium, the corresponding invasion reproduction number is given by:

$$R_{\text{inv}}(\nu) \;=\; \rho\Big(K^{\text{eff}}_{\text{inv}}(\nu)\Big)\,. \tag{11.37}$$

Explicitely, we can write:

$$\begin{aligned} R_{(SI\,|\,IS+II)} &= \rho\Big(D\Big(\mathbf{S}^*_{(IS+II)}\Big)\; K^{SI}_{\text{inv}}(\nu)\Big)\,, \\ R_{(IS\,|\,SI+II)} &= \rho\Big(D\Big(\mathbf{S}^*_{(SI+II)}\Big)\; K^{IS}_{\text{inv}}(\nu)\Big)\,, \end{aligned} \tag{11.38}$$

These expressions are formally identical to the invasion criteria in two-pathogen systems without co-infection. The novelty is entirely embedded in $K_{\text{inv}}(\nu)$, whose blocks depend on ν through the resident equilibrium $(IS{+}II)$ or $(SI + II)$ and through the transitions leading from single infection to co-infection. In particular, ν enters the invasion operator both through the probabilities of progression from single infection to co-infection (e.g. $P^{SI\to II}$ and $P^{IS\to II}$, which scale as $\tau(\lambda)$ and $\tau_q(\lambda)$), and through the resident equilibrium itself: the background fields $(\rho^{IS*}, \rho^{II*}, \rho^{SS*})$ depend on ν, thereby modifying the epidemiological background in which the invading strain grows.

11.5 Results

In this section, we analyze the dynamical behaviour of the SIIS metapopulation model under a range of epidemiological and mobility conditions. Our goal is to characterise how transmission probabilities, mobility patterns, interaction costs and coinfection penalties shape the qualitative outcomes of

competition between the two strains, and to assess the consistency between deterministic (Markovian) simulations, stochastic agent-based Monte Carlo simulations, and the analytical predictions derived from the reproduction numbers R_0 and R_{inv}.

First, we explore the dynamical regimes emerging from the discrete-time Markov equations under systematic variation of the relevant parameters. Second, we compare these results with stochastic simulations, highlighting common features and divergences, especially in the presence of coinfection. Finally, we focus on the invasion reproduction number R_{inv} and show how it provides a mathematically precise prediction of a sharp threshold in mobility, denoted $p^\star$, separating dominance of one strain from dominance of the other.

11.5.1 Deterministic Markov simulations

We begin by analyzing the deterministic behaviour of the SIIS metapopulation model through the numerical integration of the discrete-time equations presented in Section 11.1.1. For each parameter configuration, the system is iterated until convergence, and the steady-state prevalence of each compartment is computed as a function of the infection rate λ and the mobility parameter p. These two-dimensional heatmaps are arranged over grids indexed by the fitness cost q and the coinfection penalty ν, thereby revealing the qualitative structure of the model's long-term dynamics.

Since the goal of this section is to characterise competition between the two strains, we focus on the IS and SI compartments.

Asymmetric mobility: IS dominance (case $\alpha = 1$, $\beta = 0$)

When IS individuals move freely and SI individuals cannot move, the IS heatmaps exhibit a uniform structure across the entire (q, ν) parameter space. In all scenarios, the sensitive strain maintains high prevalence without any significant transitions or competing regimes.

Figure 11.2 shows a selection of heatmaps for this configuration. The smooth gradients in λ and p, combined with the absence of sharp boundaries, confirm that SI is unable to invade under any considered parameter values. This is consistent with the dual disadvantage experienced by SI: reduced transmissibility ($q < 1$) and complete loss of mobility ($\beta = 0$).

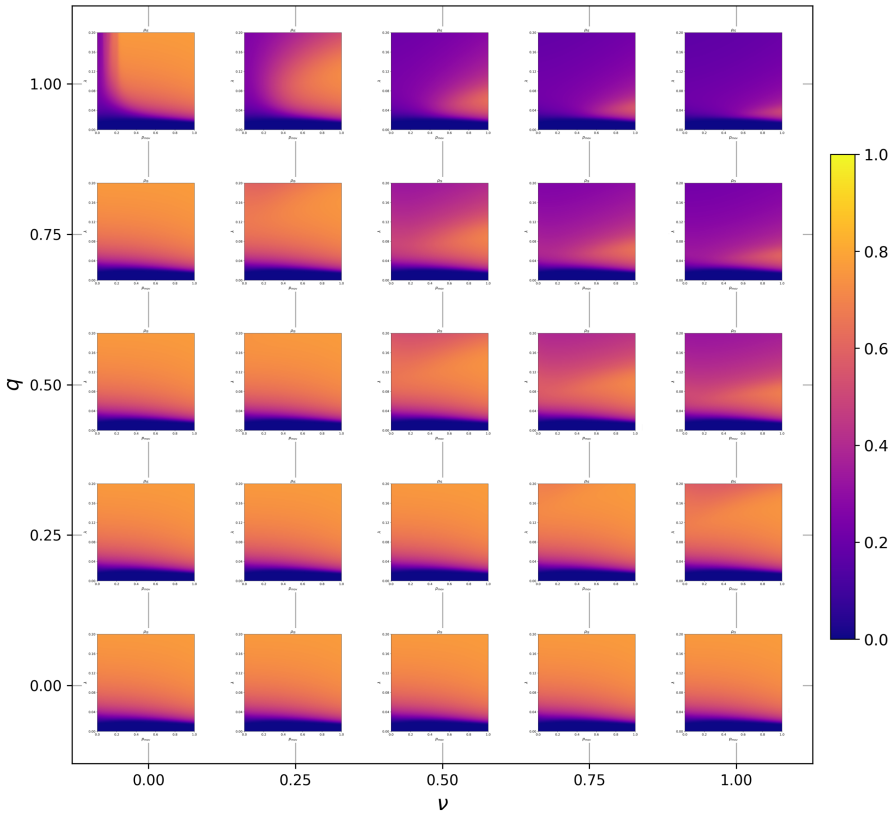


Figure 11.2: **Steady-state prevalence of IS for $\alpha = 1$, $\beta = 0$ across the (q, ν) grid.** Each panel shows the heatmap over (p, λ). No transition or SI invasion occurs in this mobility regime.

Asymmetric mobility: mobility-driven threshold (case $\alpha = 0$, $\beta = 1$)

When mobility asymmetry is reversed, IS individuals cannot move while SI individuals move freely. In this regime, for $\nu = 0$ the qualitative behaviour changes drastically. The IS heatmaps now display a sharp, nearly vertical transition in the (p, λ) plane: for low mobility, IS persists; beyond a critical threshold $p^\star$, IS collapses abruptly.

Fig. 11.3 shows the full grid of IS heatmaps for this configuration, while Fig. 11.4 shows the complementary SI heatmaps. The two sets of transitions are perfectly symmetric: SI remains absent for $p < p^\star$ but rapidly dominates for $p > p^\star$. This structure is characteristic of pure competitive exclusion with

no possibility of coexistence when $\nu = 0$.

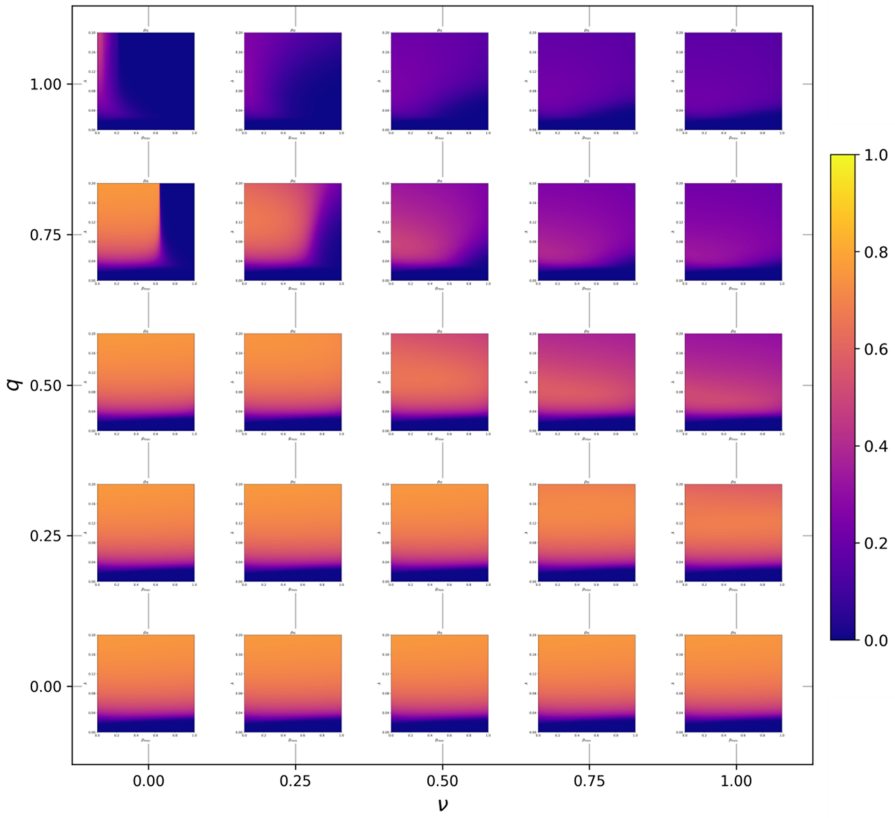


Figure 11.3: **Steady-state prevalence of IS for $\alpha = 0$, $\beta = 1$ across the (q, ν) grid.** A sharp transition at mobility $p^\star$ marks the collapse of IS for $\nu = 0$.

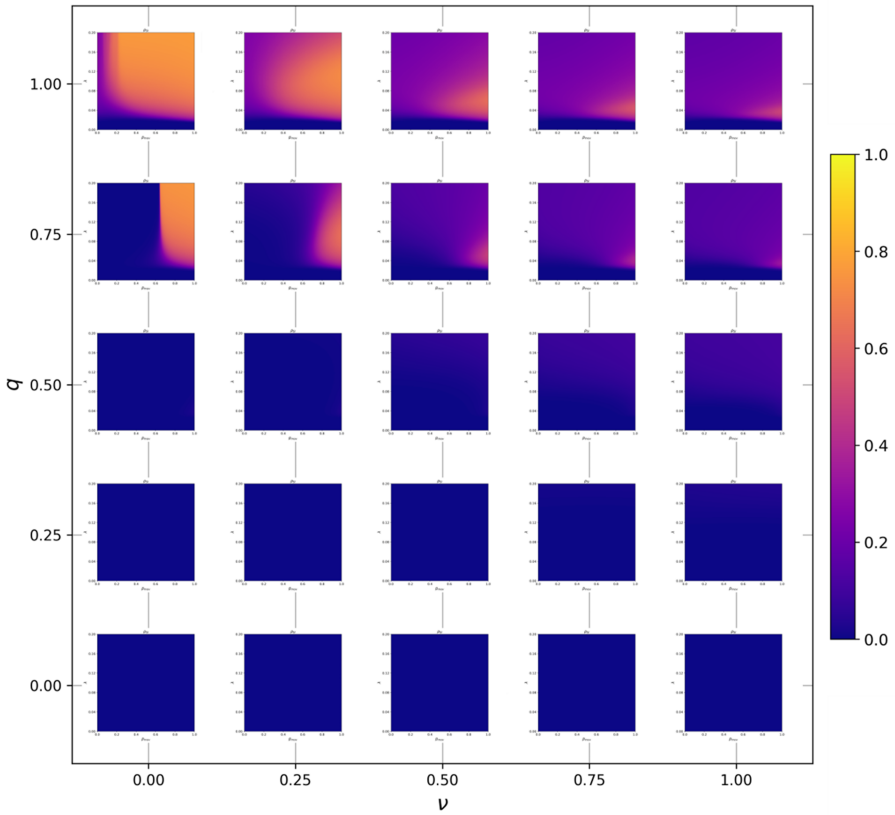


Figure 11.4: **Steady-state prevalence of SI for $\alpha = 0$, $\beta = 1$ across the (q, ν) grid.** SI becomes dominant for $p > p^\star$, mirroring the collapse of IS for $\nu = 0$.

To provide a clearer illustration, Fig. 11.5 shows the individual IS and SI heatmaps for $q = 0.80$ and $\nu = 0$.

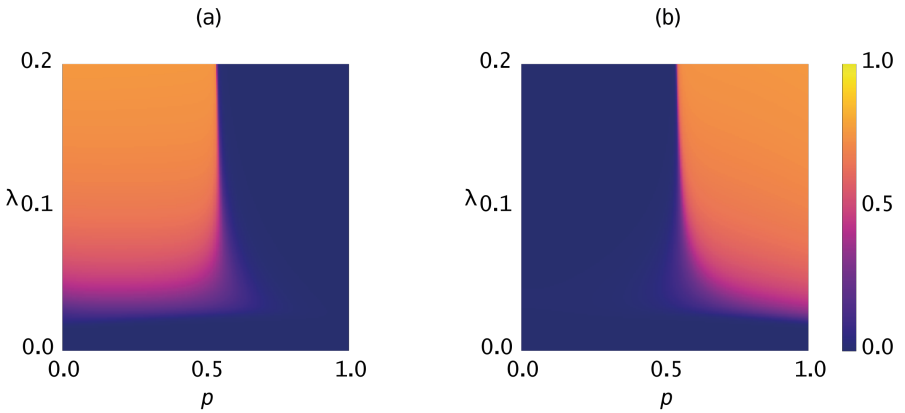


Figure 11.5: **Stationary densities** ρ_{IS} (a) and ρ_{SI} (b) for $q = 0.8$, $\nu = 0.3$ and asymmetric mobility ($\alpha = 0,\ \beta = 1$). The collapse of IS occurs at $p \approx p^\star$, when SI becomes dominant.

This pair of heatmaps makes the competitive dynamics explicit: below the mobility threshold, spatial confinement allows IS to remain endemic; above it, the resistant strain fully invades despite its intrinsic fitness cost. No coexistence region is observed, confirming the competitive-exclusion nature of the $\nu = 0$ dynamics.

The deterministic simulations highlight some structural features of the SIIS dynamics. First, mobility asymmetry emerges as the principal determinant of competitive outcomes. When only the sensitive strain is allowed to move, the resistant strain systematically fails to invade; conversely, when mobility is granted exclusively to the resistant strain, the sensitive strain collapses as soon as the mobility parameter exceeds a critical value $p^\star$. In both cases, it is the differential access to movement, and not the intrinsic transmissibility parameters, that determines which strain colonises the metapopulation. A second observation concerns the presence of a well-defined mobility threshold $p^\star$, sharply separating the IS-dominated and SI-dominated regimes, when the parameter ν is zero. This transition is rapid and remains clearly identifiable across a wide range of q values. As shown in next Subsection 11.5.3,

the analytically computed value of $p^\star$, obtained by solving the invasion condition $R_{\text{inv}} = 1$, coincides with the location of the transition observed in the deterministic simulations.

11.5.2 Monte Carlo agent-based simulations

To complement the deterministic analysis based on the discrete-time Markov equations, we performed a set of Monte Carlo (MC) agent-based simulations on the same metapopulation structure used throughout this chapter. The purpose of this section is to assess the robustness of the deterministic predictions when stochastic effects and finite-size fluctuations are explicitly accounted for, and to verify to what extent the sharp transition between dominance of the IS and SI compartments, observed in the Markov framework for $\nu = 0$, persists in a fully stochastic scenario.

Simulation setup

The MC simulations track the individual-level dynamics of 215 640 agents distributed across the 22 nodes of the Cali mobility network, with initial populations matching the empirical node sizes.

At each time step (out of a total of 500) the MC dynamics proceeds through two sequential phases. In the first phase (mobility), each agent decides whether to leave its current node. Individuals in state SS move with probability p, those in IS with probability αp, those in SI with probability βp, and agents in state II with probability $\alpha\beta p$. Whenever movement occurs, the destination node is selected according to the mobility weights encoded in the matrix R. In the second phase (infection and recovery), each agent interacts with a random fraction z of the individuals present in its destination node. Transitions between epidemiological states follow the same probabilistic rules used in the Markov model. An SS agent becomes infected by the sensitive strain with probability λ upon contact with an IS or II individual, and infected by the resistant strain with probability $q\lambda$ upon contact with an SI or II individual. Agents in IS (respectively SI) recover with probability μ; if recovery does not occur, they may acquire the second strain upon contact with SI or II (respectively IS or II) with probabilities modulated by the parameter ν. Finally, agents in state II recover from one

of the two strains with probability μ, transitioning back to either IS or SI.

For each parameter set, we ran 50 independent realisations, and results were averaged across runs. This averaging procedure is sufficient for convergence, given the large agent population, although some residual stochastic variability remains visible in the time-series plots.

Heatmap analysis for $\nu = 0$

We begin by analyzing the case $\alpha = 0$, $\beta = 1$, $q = 0.8$, and $\nu = 0$, which represents the exact stochastic analogue of the deterministic scenario shown in Section 11.5.1. In this regime, co-infection is suppressed, and the system reduces to a competition between IS and SI, modulated by mobility and by the fitness cost q. Figure 11.6 reports the heatmaps of the stationary densities ρ_{IS} and ρ_{SI} obtained from the MC simulations.

In the agent-based simulations, the densities ρ_{IS} and ρ_{SI} are obtained by counting how many agents occupy each epidemiological state. For each patch i, if $n_i^A(t)$ denotes the number of agents in state $A \in \{SS, IS, SI, II\}$ at time t and n_i the total population of that patch, the global fraction in state A is computed simply as:

$$\rho^A(t) = \frac{1}{n} \sum_{i=1}^{N} n_i^A(t),$$

where n is the total number of agents in the system and N is the number of nodes. Because single realisations are affected by stochastic fluctuations, the stationary density reported in the heatmaps is the average over the $R = 50$ independent runs. If $\rho^{A(r)}$ denotes the stationary density from run r (obtained by time-averaging $\rho^A(t)$ over the final portion of the simulation), the value shown is:

$$\bar{\rho}^A = \frac{1}{R} \sum_{r=1}^{R} \rho^{A(r)}.$$

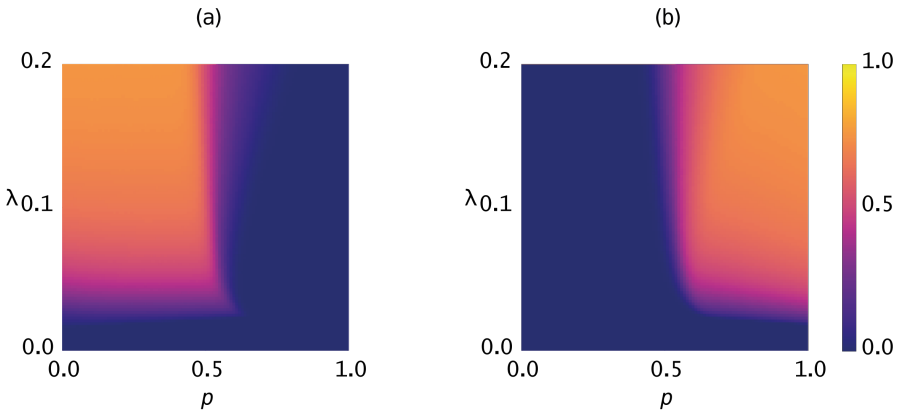


Figure 11.6: **Monte Carlo stationary densities** ρ_{IS} (a) and ρ_{SI} (b) for $q = 0.8$, $\nu = 0.3$ and asymmetric mobility ($\alpha = 0,\ \beta = 1$)

The stochastic heatmaps reproduce the key qualitative feature already observed in the Markov model: a sharp mobility-driven transition separating a region where IS dominates (p below a critical threshold $p^\star$) from a region where SI becomes the prevalent infection. The threshold is slightly blurred by stochastic effects, but its location coincides remarkably well with the deterministic prediction.

Effect of co-infection: comparison between $\nu = 0$ and $\nu > 0$

To evaluate the impact of co-infection on the dynamics, we performed additional MC simulations for the same parameter set ($\alpha = 0$, $\beta = 1$, $q = 0.8$) but with $\nu = 0.3$. Figure 11.7 shows the resulting heatmaps.

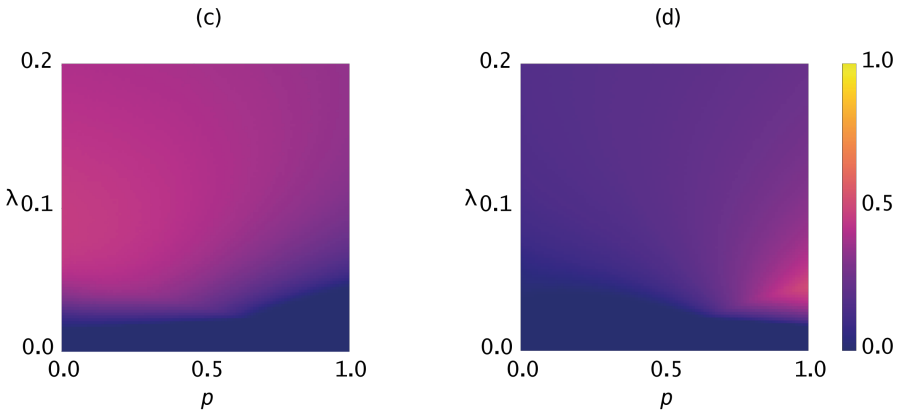


Figure 11.7: **Monte Carlo stationary densities** ρ_{IS} (a) and ρ_{SI} (b) for $q = 0.8$, $\nu = 0.3$ and asymmetric mobility ($\alpha = 0,\ \beta = 1$)

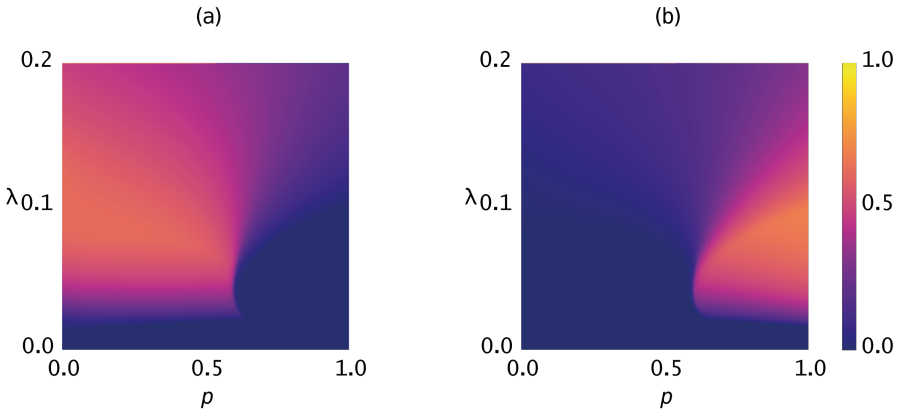


Figure 11.8: **Markov stationary densities** ρ_{IS} (a) and ρ_{SI} (b) for $q = 0.8$, $\nu = 0.3$ and asymmetric mobility ($\alpha = 0,\ \beta = 1$)

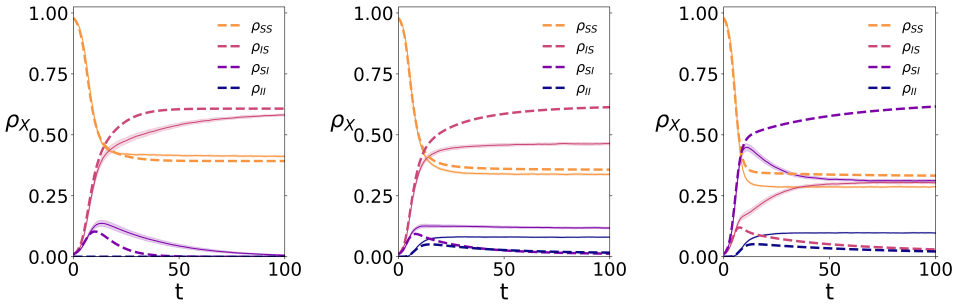
In this case, the deterministic and stochastic models diverge more noticeably. Two mechanisms offer a plausible interpretation of this discrepancy. First, when $\nu > 0$ the transitions involving acquisition of the second strain (IS $\to$ II and SI $\to$ II) become comparatively rare events. In the deterministic Markov equations, these transitions are treated as smooth,

continuous fluxes, whereas in the agent-based system, they occur through discrete, stochastic encounters. Second, the presence of the II compartment introduces an additional source of variability: once agents enter II, the sequence and timing of partial recoveries ($\mathrm{II} \to \mathrm{IS}$ or $\mathrm{II} \to \mathrm{SI}$) depend strongly on chance realizations. These stochastic asymmetries can momentarily favour one strain over the other, effectively smoothing the competition boundary observed in the $\nu = 0$ case. Even at the qualitative level, the Markov predictions (shown in Fig. 11.8) display a more structured transition than the MC simulations, which tend to smooth out the competition boundary and attenuate the sharp dominance observed at $\nu = 0$.

The MC simulations confirm the deterministic results for $\nu = 0$, while revealing the limits of the Markov approximation once co-infection becomes sufficiently frequent.

Temporal dynamics

To further compare deterministic and stochastic dynamics, we analyzed the temporal evolution of the system for three representative parameter sets. Figure 11.9 displays the average MC trajectories (thick lines with shaded variability bands) overlaid with the corresponding Markov solutions (dashed lines).



(a) (b) (c)

Figure 11.9: **Temporal evolution of the SIIS model under three representative parameter sets:** (a) $\alpha = 0$, $\beta = 1$, $q = 0.8$, $\nu = 0$, $\lambda = 0.075$, $p = 0.2$; (b) $\alpha = 0$, $\beta = 1$, $q = 0.8$, $\nu = 0.3$, $\lambda = 0.09$, $p = 0$; (c) $\alpha = 0$, $\beta = 1$, $q = 0.8$, $\nu = 0.3$, $\lambda = 0.09$, $p = 0.9$.

For $\nu = 0$, the agreement between MC and Markov trajectories is very

good, in particular at stationarity. When $\nu > 0$, however, the MC simulations systematically drift away from the deterministic prediction. This behaviour is consistent with the heatmap results: co-infection amplifies stochastic fluctuations and weakens the deterministic separation between the compartments, leading to broader stationary distributions.

11.5.3 Threshold analysis via the invasion reproduction number

The theoretical framework introduced in Section 9.5.1 provides a general criterion to analyze competitive interactions between strains in structured systems. In the SIIS model, the basic reproduction number R_0 is not informative for determining which strain dominates, since for the relevant parameter ranges of interest, we obtain $R_0 > 1$ for both IS and SI. In this regime, the classical threshold interpretation of R_0 breaks down: both strains are capable of persisting in isolation, and the outcome of competition is determined not by emergence but by *invasion.* For this reason, the appropriate indicator is the invasion reproduction number R_{inv}, which quantifies whether one strain can invade when the other is already endemic.

As discussed in Sections 9.5.1 and in 11.3, R_{inv} is computed by linearising the dynamics of the *invading* strain around the endemic equilibrium of the *resident* strain. Invasion succeeds if $R_{\text{inv}} > 1$ and fails if $R_{\text{inv}} < 1$. When $R_{\text{inv}} = 1$, the system lies precisely on an invasion threshold.

In particular, for our SIIS model, we focus on the regime with $\nu = 0$, where coexistence is structurally precluded by the absence of co-infection. In this case, the system necessarily converges to either an IS-dominated or SI-dominated state, and the transition between these two outcomes is governed by a single critical mobility value $p^\star$. To determine this threshold analytically, we compute the value of p such that the invasion criterion $R_{\text{inv}} = 1$ is satisfied. The calculation is performed for fixed

$$\lambda = 0.16, \quad \mu = 0.25, \quad \beta = 1, \quad z = 0.01, \quad \nu = 0,$$

while varying between 0 and 1, the pair (α, q), which respectively control mobility suppression and relative transmissibility of the IS strain.

First, we calculated the threshold value $p^\star$ from the model simulations with the Markov equations. In this case, $p^\star$ corresponds to the value of p at

which the transition from IS-type infected to SI-type infected occurs. Figure 11.10 reports the critical mobility values $p^\star$ obtained from deterministic Markov simulations. White regions correspond to parameter combinations for which IS always dominates (i.e., SI never invades), and no threshold can be defined.

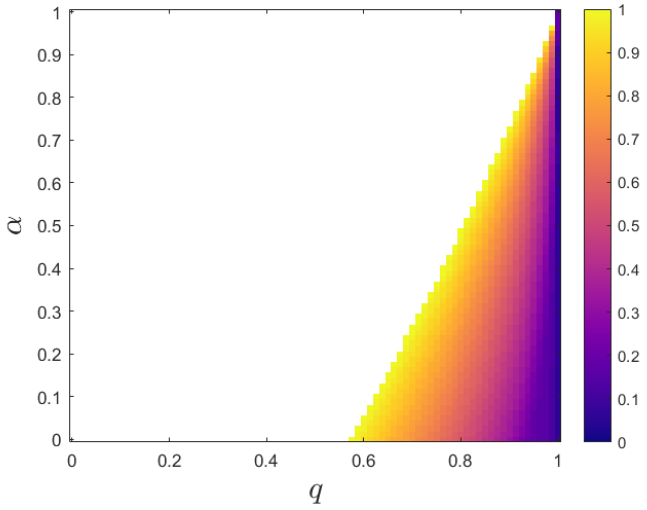


Figure 11.10: **Critical mobility values $p^\star$ estimated from deterministic Markov simulations, as a function of (α, q).** White regions indicate absence of an SI-to-IS transition.

The structure is monotonic: increasing q (i.e., increasing the effective transmissibility of SI relative to IS) shifts the threshold toward lower values of p, as expected. Similarly, increasing α increases the mobility of IS (and therefore favors IS), raising the threshold relative to $p^\star$.

Then we calculated the value of $p^\star$ for the parameter pairs (α, q) using the mathematical formulation of R_{inv}. Figure 11.11 displays the values of $p^\star$ obtained by solving the condition $R_{\text{inv}} = 1$ under the same parameter configuration.

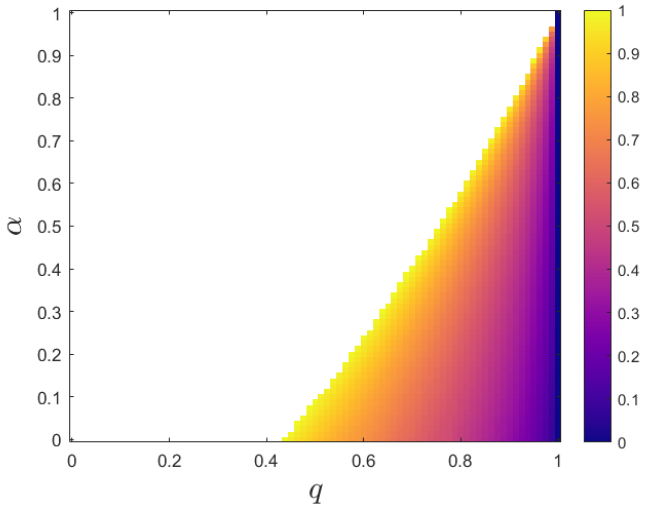


Figure 11.11: **Critical mobility values $p^\star$ obtained by solving $R_{\text{inv}} = 1$.** The white regions correspond to the parameter pairs for which SI invasion never occurs.

The agreement with the simulation-based heatmap is remarkably good. Although the analytical threshold tends to slightly underestimate $p^\star$ in regions where the deterministic dynamics exhibit a very sharp transition, the qualitative structure is preserved and the quantitative discrepancy remains small. This confirms that the invasion reproduction number captures the fundamental mechanism underlying strain replacement in the SIIS model.

Furthermore, the consistency between Fig. 11.10 and 11.11 has two main implications: the competition between IS and SI in the absence of co-infection is governed by an invasion threshold, and the sign of $R_{\text{inv}} - 1$ provides a reliable predictor of the long-term outcome, even in a structured metapopulation model with mobility heterogeneities.

For completeness, Table 11.1 summarises the relevant entries of the general classification of multi-strain invasion scenarios. The SIIS model with $\nu = 0$ fits perfectly into the regime

$$R_{0,1} > 1, \quad R_{0,2} > 1, \quad (R_{12} > 1 > R_{21}) \text{ or } (R_{21} > 1 > R_{12}),$$

corresponding to competitive exclusion: whichever strain can invade the endemic equilibrium of the other ultimately dominates.

Conditions	**Interpretation**
$R_{0,1} < 1$, $R_{0,2} < 1$	Both strains go extinct
$R_{0,1} > 1$, $R_{0,2} < 1$	Strain 1 spreads (and vice versa)
$R_{0,1} > 1$, $R_{0,2} > 1$, $R_{12} > 1 > R_{21}$	Strain 1 excludes strain 2
$R_{0,1} > 1$, $R_{0,2} > 1$, $R_{21} > 1 > R_{12}$	Strain 2 excludes strain 1
$R_{0,1} > 1$, $R_{0,2} > 1$, $R_{12} > 1$, $R_{21} > 1$	Possible coexistence / bistability

Table 11.1: Relevant invasion regimes for two competing strains. The SIIS model with $\nu = 0$ belongs to the highlighted regime.

Chapter 12

Conclusion

AMR remains one of the most challenging problems in contemporary healthcare, driven by the interplay between biological evolution, clinical practices, and the structural organization of hospitals and communities. Its persistence reflects the difficulty of capturing, within a single analytical framework, both the complexity of microbial data and the dynamical mechanisms that govern the spread and replacement of resistant strains.

This thesis approached the problem from two complementary perspectives. The first part focused on data-driven methodologies, investigating how ML techniques can be used to predict resistance phenotypes and reconstruct incomplete AMR datasets. The second part adopted a mechanistic point of view, developing and analysing mathematical models that describe the transmission of sensitive and resistant strains in structured populations, both at the ward level and across larger mobility networks.

The findings from Part I show that ML can extract predictive structure from genomic and phenotypic data despite noise, imbalance, and missingness, conditions that are typical of real AMR surveillance systems. Accurate phenotype prediction demonstrates that resistance determinants often leave identifiable signatures in genomic profiles, whereas the success of imputation methods highlights the possibility of partially mitigating the incompleteness of current datasets. At the same time, the analysis makes clear that ML models are sensitive to the biological relevance and representativeness of the available features: when determinants of resistance are weakly expressed in the data or missing entirely, predictive reliability deteriorates. These results

reinforce the idea that ML is a valuable component of AMR analysis, but one that must operate in conjunction with domain knowledge and mechanistic insight.

Part II examined AMR from a population-dynamical perspective. The XSR and SIIS models developed in this thesis reveal how spatial heterogeneity, selective pressure, fitness costs, and mobility patterns shape the competition between sensitive and resistant strains. The analytical derivation of reproduction numbers and invasion thresholds provides explicit conditions under which resistance can emerge, persist, or be outcompeted, conditions that depend not only on biological traits but also on the structure of hospital wards and, in the case of the SIIS model, on mobility across urban districts. These models show that resistance dynamics cannot be understood purely from static data, but are the result of interactions between local transmission, patient movement, and intervention strategies. In this sense, the models offer mechanistic explanations for patterns that may be observed empirically but are not directly inferable from data alone.

Furthermore, the introduction of the SIIS model represents an original element of this thesis. By incorporating both spatial mobility and competitive interactions between strains, it offers a flexible modelling framework that extends beyond AMR and can be applied to a variety of multi-strain epidemiological contexts.

The two parts of the thesis point toward an integrative framework in which ML and mechanistic modelling reinforce each other. ML-derived predictions and imputed datasets can reduce uncertainty in parameter estimation, inform the structure of transmission models, and identify heterogeneities that require explicit representation. Conversely, mechanistic models can guide the interpretation of ML outputs, highlight parameter regimes where data-driven methods are likely to fail, and provide a theoretical basis for designing new data collection strategies. This interplay suggests that neither approach is sufficient on its own: the empirical breadth of ML and the interpretative power of modelling must coexist to produce analyses that are both accurate and epidemiologically meaningful.

The work also highlights several open challenges. On the data-driven side, the main limitation remains the scarcity of large, standardised, biologically rich datasets. Expanding genomic and phenotypic surveillance and

incorporating measures of uncertainty into ML predictions will be crucial for improving robustness. From the modelling perspective, the assumptions of homogeneous mixing, deterministic updates, and simplified mobility may overlook important sources of variability, calling for models that incorporate stochasticity, behavioural heterogeneity, and more realistic data sources. A deeper integration of the two approaches will require hybrid frameworks capable of learning from data while respecting mechanistic constraints, potentially through data assimilation, Bayesian inference, or ML-guided parameter estimation.

In conclusion, this thesis contributes both methodological tools and conceptual insight into the analysis of AMR. It shows that combining data-driven and mechanistic perspectives is not merely advantageous but necessary to understand the multiscale processes that govern resistance dynamics. In this sense, this work provides a foundation for future developments, encouraging the design of analytical frameworks that are not only more predictive but also more interpretable and better aligned with the practical needs of clinicians and public health practitioners, with the ultimate goal of supporting more effective interventions against one of the most pressing challenges of modern medicine.

Appendix A

Mathematical Derivations for the XSR Metapopulation Model

This appendix provides the algebraic details omitted from Chapter 10 for readability. All derivations are consistent with the notation introduced therein. We report: (i) the explicit expression of the fourth equilibrium point of the one-node XSR system; (ii) the full linearisation of the infection probabilities; (iii) the detailed construction of the transmission and transition matrices for the NGM; (iv) the derivation connecting the global reproduction number R_0 to the critical transmission threshold β_1.

A.1 Fixed points of the one-node XSR system

The discrete-time system describing the dynamics of the three-compartment XSR model is:

$$\begin{aligned} S(t+1) &= S(t) + \beta_1 S(t)X(t) - (\gamma + \tau_1 + \tau_2)S(t) - \sigma S(t), \\ R(t+1) &= R(t) + \beta_2 R(t)X(t) + \sigma S(t) - (\gamma + \tau_2)R(t), \\ X(t) &= 1 - S(t) - R(t), \end{aligned} \tag{A.1}$$

with $\beta_2 = \beta_1(1-c)$.

At equilibrium we set $S(t+1) = S(t) = S$ and $R(t+1) = R(t) = R$,

which yields:

$$\begin{aligned} S\left[\beta_1(1-S-R)-(\gamma+\tau_1+\tau_2+\sigma)\right]=0,\\ R\left[\beta_2(1-S-R)-(\gamma+\tau_2)\right]+\sigma S=0. \end{aligned} \tag{A.2}$$

The first three equilibrium points follow immediately:

$$(S^*,R^*)=(0,0), \qquad \left(1-\frac{\gamma+\tau_1+\tau_2+\sigma}{\beta_1},\,0\right), \qquad \left(0,\,1-\frac{\gamma+\tau_2}{\beta_2}\right).$$

The fourth equilibrium point, corresponding to the coexistence of S and R, is obtained by solving system (A.2) under the condition $S \neq 0$ and $R \neq 0$. The explicit closed-form expression is:

$$\begin{aligned} S^* &= \frac{(-\beta_1+\gamma+\tau_1+\tau_2+\sigma)\left[-\beta_1(\gamma+\tau_2)+\beta_2(\gamma+\tau_1+\tau_2+\sigma)\right]}{\beta_1\left[-\beta_1(\gamma+\tau_2+\sigma)+\beta_2(\gamma+\tau_1+\tau_2+\sigma)\right]},\\ R^* &= \frac{\sigma(-\beta_1+\gamma+\tau_1+\tau_2+\sigma)}{-\beta_1(\gamma+\tau_2+\sigma)+\beta_2(\gamma+\tau_1+\tau_2+\sigma)}. \end{aligned} \tag{A.3}$$

As discussed in Chapter 10, when $\sigma \approx 0$ coexistence occurs only in the degenerate case:

$$\frac{\gamma+\tau_1+\tau_2}{\beta_1}=\frac{\gamma+\tau_2}{\beta_2}.$$

A.2 Linearisation of infection probabilities

The infection events depend on the probabilities:

$$P_i^S = 1-\prod_{j=1}^{N}(1-\beta_1\rho_j^S)^{n_{j\to i}}, \qquad P_i^R = 1-\prod_{j=1}^{N}(1-\beta_2\rho_j^R)^{n_{j\to i}}, \tag{A.4}$$

where $n_{j\to i}$ is defined as in Eq. (2.13) of the main text.

Near the disease-free equilibrium, $\rho_j^S, \rho_j^R \ll 1$ and we may use:

$$(1-x)^k \approx 1-kx \qquad \text{for small } x.$$

Thus:

$$P_i^S \approx \sum_{j=1}^{N}\beta_1\,\rho_j^S\,n_{j\to i}, \qquad P_i^R \approx \sum_{j=1}^{N}\beta_2\,\rho_j^R\,n_{j\to i}.$$

Substituting into the mobility-weighted transition terms:

$$\Pi_i^{X\to S} = (1-p_d)P_i^S + p_d\sum_{j=1}^{N}R_{ij}P_j^S,$$

we obtain:

$$\Pi_i^{X\to S} \approx (1-p_d)\sum_{j=1}^{N}\beta_1\rho_j^S n_{j\to i} + p_d\sum_{j=1}^{N}R_{ij}\left(\sum_{l=1}^{N}\beta_1\rho_l^S n_{l\to j}\right). \qquad \text{(A.5)}$$

The same calculation applies to $\Pi_i^{X\to R}$.

To make the structure explicit, we substitute $n_{j\to i}$:

$$n_{j\to i} = \delta_{ij}(1-p_d)n_i + p_d\,R_{ji}n_j,$$

which yields:

$$\Pi_i^{X\to S} \approx \beta_1\sum_{j=1}^{N}\Big[(1-p_d)^2\delta_{ij}n_j + p_d(1-p_d)\,n_j\,(R_{ij}+R_{ji}) + p_d^2\,n_j(RR^T)_{ij}\Big]\,\rho_j^S. \qquad \text{(A.6)}$$

This motivates the definition of the matrix M:

$$M_{ij} = (1-p_d)^2\,\delta_{ij}n_j + p_d(1-p_d)\,n_j(R_{ij}+R_{ji}) + p_d^2\,n_j(RR^T)_{ij}. \qquad \text{(A.7)}$$

Thus:

$$\Pi_i^{X\to S} \approx \beta_1(M\vec{\rho}^S)_i, \qquad \Pi_i^{X\to R} \approx \beta_2(M\vec{\rho}^R)_i.$$

A.3 Derivation of the NGM matrices

The linearised subsystem for the infected variables at the disease-free equilibrium may be written as:

$$\mathbf{X}(t+1) = (T+\Sigma)\mathbf{X}(t).$$

From Eqs. (3.12)–(3.15) we obtain the local blocks:

$$T_i = \begin{pmatrix}\beta_1 n_i & 0\\ 0 & \beta_2 n_i\end{pmatrix}, \qquad \Sigma_i = \begin{pmatrix}1-(\gamma+\tau_1+\tau_2+\sigma) & 0\\ \sigma & 1-(\gamma+\tau_2)\end{pmatrix}.$$

The global matrices are block-diagonal for Σ:

$$\Sigma = \begin{pmatrix}\mathcal{I}_N(1-(\gamma+\tau_1+\tau_2+\sigma)) & \mathcal{O}_N\\ \mathcal{I}_N(\sigma) & \mathcal{I}_N(1-(\gamma+\tau_2))\end{pmatrix},$$

and block-structured for T:

$$T = \begin{pmatrix} \left[(1-p_d)\beta_1 n_{k\to i} + p_d\beta_1 R_{ik} n_k^{\text{eff}}\right] & \mathcal{O}_N \\ \mathcal{O}_N & \left[(1-p_d)\beta_2 n_{k\to i} + p_d\beta_2 R_{ik} n_k^{\text{eff}}\right] \end{pmatrix}.$$

The inverse needed for the NGM is:

$$(\mathcal{I}-\Sigma)^{-1} = \begin{pmatrix} \mathcal{I}_N \frac{1}{\gamma+\tau_1+\tau_2+\sigma} & \mathcal{O}_N \\ \mathcal{I}_N \frac{\sigma}{(\gamma+\tau_2)(\gamma+\tau_1+\tau_2+\sigma)} & \mathcal{I}_N \frac{1}{\gamma+\tau_2} \end{pmatrix}.$$

Multiplying $T(\mathcal{I}-\Sigma)^{-1}$ yields the block-lower-triangular NGM:

$$K = \begin{pmatrix} K_{11} & \mathcal{O}_N \\ K_{21} & K_{22} \end{pmatrix},$$

with:

$$K_{11,ij} = \frac{\beta_1}{\gamma+\tau_1+\tau_2+\sigma}\left[(1-p_d)n_{j\to i} + p_d R_{ij} n_j^{\text{eff}}\right],$$
$$K_{22,ij} = \frac{\beta_2}{\gamma+\tau_2}\left[(1-p_d)n_{j\to i} + p_d R_{ij} n_j^{\text{eff}}\right].$$

A.4 Equivalence between R_0 and the critical transmission threshold

Let $\lambda_{\max}(M)$ be the spectral radius of the matrix M defined in Eq. (A.7). From the NGM structure:

$$R_0 = \max\left\{\lambda_{\max}(K_{11}), \lambda_{\max}(K_{22})\right\}.$$

Since K_{11} and K_{22} are proportional to M:

$$\lambda_{\max}(K_{11}) = \frac{\beta_1}{\gamma+\tau_1+\tau_2+\sigma}\lambda_{\max}(M),$$
$$\lambda_{\max}(K_{22}) = \frac{\beta_2}{\gamma+\tau_2}\lambda_{\max}(M),$$

and recalling $\beta_2 = (1-c)\beta_1$, we obtain:

$$R_0 = \lambda_{\max}(M)\max\left\{\frac{\beta_1}{\gamma+\tau_1+\tau_2+\sigma}, \frac{(1-c)\beta_1}{\gamma+\tau_2}\right\}.$$

Setting $R_0 = 1$ and solving for β_1 yields:

$$\beta_{1c} = \frac{1}{\lambda_{\max}(M)}\min\left\{\gamma+\tau_1+\tau_2+\sigma,\ \frac{\gamma+\tau_2}{1-c}\right\},$$

which matches the expression in Chapter 10.

Publications

Condorelli, C., Nicitra, E., Musso, N., Bongiorno, D., Stefani, S., Gambuzza, L. V., Carchiolo V., Frasca, M. (2024). *Prediction of antimicrobial resistance of Klebsiella pneumoniae from genomic data through machine learning.* Plos one, 19(9), e0309333, https://doi.org/10.1371/journal.pone.0309333.

Condorelli, C., Carchiolo V., Frasca, M., Gambuzza, L. V. (2025). *Machine Learning Approaches for Handling Missing Data in Antimicrobial Resistance Databases.*in IEEE Access, vol. 14, pp. 2015-2033, 2026, doi: 10.1109/ACCESS.2025.3650085.

Condorelli, C., Gallarta-Sáenz, P., Gambuzza, L. V., Frasca, M., Gómez-Gardeñez, J. (2026). *Mobility-driven competition between two pathogen strains in spatially structured metapopulation models.* In preparation.

Bibliography

[1] A. Adadi. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1):24, 2021.

[2] L. J. Allen and P. van den Driessche. The basic reproduction number in some discrete-time epidemic models. *Journal of difference equations and applications*, 14(10-11):1127–1147, 2008.

[3] A. K. Anand Kumar, D. Roberts, K. Wood, B. Light, J. Parrillo, S. S. Satendra Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. 2006.

[4] R. M. Anderson and R. M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1991.

[5] D. I. Andersson and D. Hughes. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260–271, 2010.

[6] D. I. Andersson and B. R. Levin. The biological cost of antibiotic resistance. *Current Opinion in Microbiology*, 2(5):489–493, 1999.

[7] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6:1–15, 2018.

[8] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: what is it and how does it work?

International journal of methods in psychiatric research, 20(1):40–49, 2011.

[9] V. Ballén, Y. Gabasa, C. Ratia, R. Ortega, M. Tejero, and S. Soto. Antibiotic resistance and virulence profiles of klebsiella pneumoniae strains isolated from different clinical sources. *Frontiers in Cellular and Infection Microbiology*, 11:738223, 2021.

[10] G. E. Batista and M. C. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533, 2003.

[11] Y. Bengio, I. Goodfellow, and A. Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[12] E. BERKERY. *Missing data analysis with the Mahalanobis distance.* PhD thesis, University of Limerick, 2016.

[13] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[14] D. Bongiorno, D. A. Bivona, C. Cicino, E. M. Trecarichi, A. Russo, N. Marascio, M. L. Mezzatesta, N. Musso, G. F. Privitera, A. Quirino, et al. Omic insights into various ceftazidime-avibactam-resistant klebsiella pneumoniae isolates from two southern italian regions. *Frontiers in Cellular and Infection Microbiology*, 12:1467, 2023.

[15] F. Brauer. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2(2):113–127, 2017.

[16] M. L. Brinch, A. Palladino, J. Geurtsen, T. Van Effelterre, L. Argante, M. J. McConnell, L. Christiansen, M. A. Pihl, N. K. Lund, and T. Hald. The neglected model validation of antimicrobial resistance transmission models–a systematic review. *Antimicrobial Resistance & Infection Control*, 14(1):59, 2025.

[17] B. Cánovas-Segura, A. Morales, A. L. Martinez-Carrasco, M. Campos, J. M. Juarez, L. L. Rodríguez, and F. Palacios. Improving interpretable prediction models for antimicrobial resistance. In *2019 IEEE*

32nd International Symposium on Computer-Based Medical Systems (CBMS), pages 543–546. IEEE, 2019.

[18] C. L. Cazer, L. F. Westblade, M. S. Simon, R. Magleby, M. Castanheira, J. G. Booth, S. G. Jenkins, and Y. T. Gröhn. Analysis of multidrug resistance in staphylococcus aureus with a machine learning-generated antibiogram. *Antimicrobial Agents and Chemotherapy*, 65(4):10–1128, 2021.

[19] S. Chang, M. L. Wilson, B. Lewis, Z. Mehrab, K. K. Dudakiya, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, et al. Supporting covid-19 policy response with large-scale mobility-based modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2632–2642, 2021.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[21] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS medicine*, 4(1):e13, 2007.

[22] V. Colizza, R. Pastor-Satorras, and A. Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282, 2007.

[23] V. Colizza and A. Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Physical review letters*, 99(14):148701, 2007.

[24] S. E. Cosgrove. The relationship between antimicrobial resistance and patient outcomes: mortality, length of hospital stay, and health care costs. *Clinical Infectious Diseases*, 42(Supplement_2):S82–S89, 2006.

[25] E. Cuevas. An agent-based model to evaluate the covid-19 transmission risks in facilities. *Computers in biology and medicine*, 121:103827, 2020.

[26] P. Dadgostar. Antimicrobial resistance: implications and costs. *Infection and drug resistance*, pages 3903–3910, 2019.

[27] O. Diekmann, H. Heesterbeek, and T. Britton. *Mathematical tools for understanding infectious disease dynamics*, volume 7. Princeton University Press, 2013.

[28] O. Diekmann, J. Heesterbeek, and M. G. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the royal society interface*, 7(47):873–885, 2010.

[29] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations. *Journal of mathematical biology*, 28(4):365–382, 1990.

[30] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[31] F. Durazzi, M. D. Pezzani, F. Arieti, O. Simonetti, L. M. Canziani, E. Carrara, L. Barbato, F. Onorati, D. Remondini, and E. Tacconelli. Modelling antimicrobial resistance transmission to guide personalized antimicrobial stewardship interventions and infection control policies in healthcare setting: a pilot study. *Scientific Reports*, 13(1):15803, 2023.

[32] E. Elyan, A. Hussain, A. Sheikh, A. A. Elmanama, P. Vuttipittayamongkol, and K. Hijazi. Antimicrobial resistance and machine learning: challenges and opportunities. *IEEE Access*, 10:31561–31577, 2022.

[33] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. A survey on missing data in machine learning. *Journal of Big data*, 8:1–37, 2021.

[34] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas. Iterative robust semi-supervised missing data imputation. *IEEE Access*, 8:90555–90569, 2020.

[35] M. Feldgarden, V. Brover, D. H. Haft, A. B. Prasad, D. J. Slotta, I. Tolstoy, G. H. Tyson, S. Zhao, C.-H. Hsu, P. F. McDermott, et al. 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):10–1128, 2019.

[36] G. Feretzakis, E. Loupelis, A. Sakagianni, D. Kalles, M. Martsoukou, M. Lada, N. Skarmoutsou, C. Christopoulos, K. Valakis, A. Velentza, et al. Using machine learning techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in greece. *Antibiotics*, 9(2):50, 2020.

[37] G. French. Clinical impact and relevance of antibiotic resistance. *Advanced drug delivery reviews*, 57(10):1514–1527, 2005.

[38] P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7-9):1483–1493, 2009.

[39] T. B. Giblin, R. L. Sinkowitz-Cochran, P. L. Harris, S. Jacobs, K. Liberatore, M. A. Palfreyman, E. I. Harrison, D. M. Cardo, C. C. to Prevent Antimicrobial Resistance Team, et al. Clinicians' perceptions of the problem of antimicrobial resistance in health care facilities. *Archives of internal medicine*, 164(15):1662–1668, 2004.

[40] R. Goel, L. Bonnetain, R. Sharma, and A. Furno. Mobility-based sir model for complex networks: with case study of covid-19. *Social Network Analysis and Mining*, 11(1):105, 2021.

[41] J. Gómez-Gardenes, D. Soriano-Panos, and A. Arenas. Critical regimes driven by recurrent mobility patterns of reaction–diffusion processes in networks. *Nature Physics*, 14(4):391–395, 2018.

[42] B. Grenfell and J. Harwood. (meta) population dynamics of infectious diseases. *Trends in ecology & evolution*, 12(10):395–399, 1997.

[43] T. Haslwanter. An introduction to statistics with python. *With Applications in the Life Sciences; Springer International Publishing: Cham, Switzerland*, 2016.

[44] S. He, S. Tang, L. Rong, et al. A discrete stochastic model of the covid-19 outbreak: Forecast and control. *Math. Biosci. Eng*, 17(4):2792–2804, 2020.

[45] J. A. P. Heesterbeek. A brief history of r 0 and a recipe for its calculation. *Acta biotheoretica*, 50(3):189–204, 2002.

[46] T. L. G. Hepatology. The problem of antimicrobial resistance in chronic liver disease, 2022.

[47] H. W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[48] A. Howard, N. Reza, P. L. Green, M. Yin, E. Duffy, H. C. Mwandumba, A. Gerada, and W. Hope. Artificial intelligence and infectious diseases: tackling antimicrobial resistance, from personalised care to antibiotic discovery. *The Lancet Infectious Diseases*, 2025.

[49] T. Jesudason. Who publishes updated list of bacterial priority pathogens. *The Lancet Microbe*, 5(9), 2024.

[50] S. I. Kampezidou, A. Tikayat Ray, A. P. Bhat, O. J. Pinon Fischer, and D. N. Mavris. Fundamental components and principles of supervised machine learning workflows with numerical and categorical data. *Eng*, 5(1):384–416, 2024.

[51] H. Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406, 2013.

[52] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton university press, 2008.

[53] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[54] W. Kirch, editor. *Pearson's Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht, 2008.

[55] G. M. Knight, N. G. Davies, C. Colijn, F. Coll, T. Donker, D. R. Gifford, R. E. Glover, M. Jit, E. Klemm, S. Lehtinen, et al. Mathematical modelling for antibiotic resistance control policy: do we know enough? *BMC infectious diseases*, 19(1):1011, 2019.

[56] T. Köse, S. Özgür, E. Coşgun, A. Keskinoğlu, and P. Keskinoğlu. Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. *BioMed Research International*, 2020(1):1895076, 2020.

[57] J. Li, S. Guo, R. Ma, J. He, X. Zhang, D. Rui, Y. Ding, Y. Li, L. Jian, J. Cheng, et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, 24(1):41, 2024.

[58] J. Li, S. Xie, S. Ahmed, F. Wang, Y. Gu, C. Zhang, X. Chai, Y. Wu, J. Cai, and G. Cheng. Antimicrobial activity and resistance: influencing factors. *Frontiers in pharmacology*, 8:364, 2017.

[59] R. Li, Y. Chen, and J. H. Moore. Integration of genetic and clinical information to improve imputation of data missing from electronic health records. *Journal of the American Medical Informatics Association*, 26(10):1056–1063, 2019.

[60] Y. Li, X. Cui, X. Yang, G. Liu, and J. Zhang. Artificial intelligence in predicting pathogenic microorganisms' antimicrobial resistance: challenges, progress, and prospects. *Frontiers in Cellular and Infection Microbiology*, 14:1482186, 2024.

[61] Y. Li, Q. Zhou, Y. Fan, G. Pan, Z. Dai, and B. Lei. A novel machine learning-based imputation strategy for missing data in step-stress accelerated degradation test. *Heliyon*, 10(4), 2024.

[62] M. Lipsitch, J. Dykes, S. Johnson, E. Ades, J. King, D. Briles, and G. Carlone. Competition among streptococcus pneumoniae for intranasal colonization in a mouse model. *Vaccine*, 18(25):2895–2901, 2000.

[63] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[64] W. Liu, N. Ying, Q. Mo, S. Li, M. Shao, L. Sun, and L. Zhu. Machine learning for identifying resistance features of klebsiella pneumoniae using whole-genome sequence single nucleotide polymorphisms. *Journal of Medical Microbiology*, 70(11):001474, 2021.

[65] X. Liu, D. Cui, H. Li, Q. Wang, Z. Mao, L. Fang, N. Ren, and J. Sun. Direct medical burden of antimicrobial-resistant healthcare-associated infections: empirical evidence from china. *Journal of Hospital Infection*, 105(2):295–305, 2020.

[66] Z. Liu, D. Deng, H. Lu, J. Sun, L. Lv, S. Li, G. Peng, X. Ma, J. Li, Z. Li, et al. Evaluation of machine learning models for predicting antimicrobial resistance of actinobacillus pleuropneumoniae from whole genome sequences. *Frontiers in microbiology*, 11:48, 2020.

[67] C. Llor and L. Bjerrum. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic advances in drug safety*, 5(6):229–241, 2014.

[68] A. L. Lloyd and R. M. May. Spatial heterogeneity in epidemic models. *Journal of theoretical biology*, 179(1):1–11, 1996.

[69] H. Lucy. Embracing a one health framework to fight antimicrobial resistance, 2023.

[70] H. Lund-Palau, A. R. Turnbull, A. Bush, E. Bardin, L. Cameron, O. Soren, N. Wierre-Gore, E. W. Alton, J. G. Bundy, G. Connett, et al. Pseudomonas aeruginosa infection in cystic fibrosis: pathophysiological mechanisms and therapeutic approaches. *Expert review of respiratory medicine*, 10(6):685–697, 2016.

[71] G. A. Lyngdoh, M. Zaki, N. A. Krishnan, and S. Das. Prediction of concrete strengths enabled by missing data imputation and interpretable machine learning. *Cement and Concrete Composites*, 128:104414, 2022.

[72] N. Macesic, O. J. Bear Don't Walk IV, I. Pe'er, N. P. Tatonetti, A. Y. Peleg, and A.-C. Uhlemann. Predicting phenotypic polymyxin resistance in klebsiella pneumoniae through machine learning analysis of genomic data. *Msystems*, 5(3):10–1128, 2020.

[73] C. Mack, Z. Su, and D. Westreich. Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide. 2018.

[74] A.-P. Magiorakos, A. Srinivasan, R. B. Carey, Y. Carmeli, M. Falagas, C. Giske, S. Harbarth, J. Hindler, G. Kahlmeter, B. Olsson-Liljequist, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clinical microbiology and infection*, 18(3):268–281, 2012.

[75] F. Malizia, L. Gallo, M. Frasca, V. Latora, and G. Russo. Individual- and pair-based models of epidemic spreading: Master equations and analysis of their forecasting capabilities. *Physical Review Research*, 4(2):023145, 2022.

[76] G. Mancuso, A. Midiri, E. Gerace, and C. Biondo. Bacterial antibiotic resistance: The most critical pathogens. *Pathogens*, 10(10):1310, 2021.

[77] M. Matesanz and J. Mensa. Ceftazidime-avibactam. *Revista Española de Quimioterapia*, 34(Suppl1):38, 2021.

[78] M. M. E. Meybodi, A. R. Foroushani, M. Zolfaghari, A. Abdollahi, A. Alipour, E. Mohammadnejad, E. Z. Mehrjardi, and A. Seifi. Antimicrobial resistance pattern in healthcare-associated infections: investigation of in-hospital risk factors. *Iranian journal of microbiology*, 13(2):178, 2021.

[79] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts. Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data. *PLoS computational biology*, 14(12):e1006258, 2018.

[80] C. J. Murray, K. S. Ikuta, F. Sharara, L. Swetschinski, G. R. Aguilar, A. Gray, C. Han, C. Bisignano, P. Rao, E. Wool, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The lancet*, 399(10325):629–655, 2022.

[81] M. Naghavi, S. E. Vollset, K. S. Ikuta, L. R. Swetschinski, A. P. Gray, E. E. Wool, G. R. Aguilar, T. Mestrovic, G. Smith, C. Han, et al. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*, 404(10459):1199–1226, 2024.

[82] M. Nguyen, S. W. Long, P. F. McDermott, R. J. Olsen, R. Olson, R. L. Stevens, G. H. Tyson, S. Zhao, and J. J. Davis. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *Journal of clinical microbiology*, 57(2):10–1128, 2019.

[83] M. Oonsivilai, Y. Mo, N. Luangasanatip, Y. Lubell, T. Miliya, P. Tan, L. Loeuk, P. Turner, and B. S. Cooper. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in cambodia. *Wellcome open research*, 3, 2018.

[84] R. Paramasivam, D. R. Gopal, R. Dhandapani, R. Subbarayalu, M. P. Elangovan, B. Prabhu, V. Veerappan, A. Nandheeswaran, S. Paramasivam, and S. Muthupandian. Is amr in dairy products a threat to human health? an updated review on the origin, prevention, treatment, and economic impacts of subclinical mastitis. *Infection and Drug Resistance*, pages 155–178, 2023.

[85] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925–979, 2015.

[86] F. Pennisi, A. Pinto, G. E. Ricciardi, C. Signorelli, and V. Gianfredi. The role of artificial intelligence and machine learning models in antimicrobial stewardship in public health: a narrative review. *Antibiotics*, 14(2):134, 2025.

[87] T. M. Pham, N. Pandis, and I. R. White. Missing data: Issues, concepts, methods. In *Seminars in Orthodontics*. Elsevier, 2024.

[88] M. Pichler and F. Hartig. Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, 14(4):994–1016, 2023.

[89] P. D. Powell. Calculating determinants of block matrices. *arXiv preprint arXiv:1112.4379*, 2011.

[90] C. N. I. S. Program et al. Healthcare-associated infections and antimicrobial resistance in canadian acute care hospitals, 2016–2020. *Canada Communicable Disease Report*, 48(7-8):308, 2022.

[91] M. Prosperi, C. Boucher, J. Bian, and S. Marini. Assessing putative bias in prediction of anti-microbial resistance from real-world genotyping data under explicit causal assumptions. *Artificial intelligence in medicine*, 130:102326, 2022.

[92] D. Qin. Next-generation sequencing and its clinical application. *Cancer biology & medicine*, 16(1):4, 2019.

[93] F. Rajer and L. Sandegren. The role of antibiotic resistance genes in the fitness cost of multiresistance plasmids. *MBio*, 13(1):e03552–21, 2022.

[94] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, A.-C. Hauschild, O. Schwengers, and D. Heider. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38(2):325–334, 2022.

[95] A. K. Rodrigues, R. Ospina, and M. R. Ferreira. Adaptive kernel fuzzy clustering for missing data. *Plos one*, 16(11):e0259266, 2021.

[96] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[97] A. Sakagianni, C. Koufopoulou, G. Feretzakis, D. Kalles, V. S. Verykios, and P. Myrianthefs. Using machine learning to predict antimicrobial resistance—a literature review. *Antibiotics*, 12(3):452, 2023.

[98] M. D. Samad, S. Abrar, and N. Diawara. Missing value estimation using clustering and deep learning within multiple imputation framework. *Knowledge-based systems*, 249:108968, 2022.

[99] M. Sartelli, C. P. Marini, J. McNelis, F. Coccolini, C. Rizzo, F. M. Labricciosa, and P. Petrone. Preventing and controlling healthcare-associated infections: the first principle of every antimicrobial stewardship program in hospital settings. *Antibiotics*, 13(9):896, 2024.

[100] E. Sauerborn, N. C. Corredor, T. Reska, A. Perlas, S. Vargas da Fonseca Atum, N. Goldman, N. Wantia, C. Prazeres da Costa, E. Foster-Nyarko, and L. Urban. Detection of hidden antibiotic resistance through real-time genomics. *Nature communications*, 15(1):5494, 2024.

[101] J. L. Schafer. Analysis of incomplete multivariate data. *Chapman &Hall/CRC*, 1997.

[102] K. Seu, M.-S. Kang, and H. Lee. An intelligent missing data imputation techniques: A review. *JOIV: International Journal on Informatics Visualization*, 6(1-2):278–283, 2022.

[103] T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torné, E. Sala, P. Lió, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.

[104] D. Soriano-Paños, F. Ghanbarnejad, S. Meloni, and J. Gómez-Gardeñes. Markovian approach to tackle the interaction of simultaneous diseases. *Physical Review E*, 100(6):062308, 2019.

[105] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[106] R. Tang, R. Luo, S. Tang, H. Song, and X. Chen. Machine learning in predicting antimicrobial resistance: A systematic review and meta-analysis. *International Journal of Antimicrobial Agents*, page 106684, 2022.

[107] M. Tharmakulasingam, B. Gardner, R. La Ragione, and A. Fernando. Explainable deep learning approach for multilabel classification of antimicrobial resistance with missing labels. *IEEE Access*, 10:113073–113085, 2022.

[108] S. Valavarasu, Y. Sangu, and T. Mahapatra. Prediction of antibiotic resistance from antibiotic susceptibility testing results from surveillance data using machine learning. *Scientific Reports*, 15(1):30509, 2025.

[109] S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[110] P. Van den Driessche and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences*, 180(1-2):29–48, 2002.

[111] A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. Kraemer, J. Wu, et al. Modelling covid-19. *Nature Reviews Physics*, 2(6):279–281, 2020.

[112] K.-D. Vihta, E. Pritchard, K. B. Pouwels, S. Hopkins, R. L. Guy, K. Henderson, D. Chudasama, R. Hope, B. Muller-Pebody, A. S. Walker, et al. Predicting future hospital antimicrobial resistance prevalence using machine learning. *Communications Medicine*, 4(1):197, 2024.

[113] P. Wang and X. Chen. Three-way ensemble clustering for incomplete data. *IEEE Access*, 8:91855–91864, 2020.

[114] S. Wang, C. Zhao, Y. Yin, F. Chen, H. Chen, and H. Wang. A practical approach for predicting antimicrobial phenotype resistance in staphylococcus aureus through machine learning analysis of genome data. *Frontiers in Microbiology*, 13:841289, 2022.

[115] W. Wang, M. Baker, Y. Hu, J. Xu, D. Yang, A. Maciel-Guerra, N. Xue, H. Li, S. Yan, M. Li, et al. Whole-genome sequencing and machine

learning analysis of staphylococcus aureus from multiple heterogeneous sources in china reveals common genetic traits of antimicrobial resistance. *Msystems*, 6(3):e01185–20, 2021.

[116] C. V. Weis, C. R. Jutzeler, and K. Borgwardt. Machine learning for microbial identification and antimicrobial susceptibility testing on maldi-tof mass spectra: a systematic review. *Clinical Microbiology and Infection*, 26(10):1310–1317, 2020.

[117] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, 13(6):e1005595, 2017.

[118] R. J. Woodman and A. A. Mangoni. A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. *Aging Clinical and Experimental Research*, 35(11):2363–2397, 2023.

[119] W. Xu, Z. Pan, Y. Wu, X.-L. An, W. Wang, B. Adamovich, Y.-G. Zhu, J.-Q. Su, and Q. Huang. A database on the abundance of environmental antibiotic resistance genes. *Scientific Data*, 11(1):250, 2024.

[120] X. Xu, W. Chong, S. Li, A. Arabo, and J. Xiao. Miaec: Missing data imputation based on the evidence chain. *IEEE Access*, 6:12983–12992, 2018.

[121] J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

[122] J. Zhao, Y. Nie, S. Ni, and X. Sun. Traffic data imputation and prediction: An efficient realization of deep learning. *IEEE Access*, 8:46713–46722, 2020.

[123] Z.-H. Zhou. *Machine learning*. Springer Nature, 2021.