Social Determinants of Methicillin-Resistant *Staphylococcus aureus* Bloodstream Infection Patterns in California

Hannah Bower[*]
hbower6@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

AJ Subudhi
asubudhi6@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Alan Wang
awang450@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

Methicillin-resistant *Staphylococcus aureus* (MRSA) remains one of the most serious antimicrobial threats in the United States, causing significant clinical, economic, and public health challenges. While the biological and clinical factors MRSA infections are well studied, the influence of social determinants of health (SDOH) such as income, housing, healthcare access, and population density, on infection patterns remains less understood. This project aims to investigate how socioeconomic factors contribute to geographic and temporal variations in MRSA infection rates across California counties over the past decade. By integrating hospital-reported infection data with county-level demographic features, this project will explore these relationships through a combination of statistical and machine learning approaches, including spatial deep learning and regression, time-series forecasting, and ensemble modeling. Analysis throughout this project will identify the most influential predictors of infection risk, provide county-level forecasts, indicate spatiotemporal impacts on risk through geographical visualizations, and highlight key areas for targeted public health interventions. The findings will uncover insights into how non-clinical factors shape MRSA epidemiology and support data-driven strategies to reduce infection burden in vulnerable communities.

1 Literature Review

Background and Significance. Methicillin-resistant *Staphylococcus aureus* (MRSA) is one of the most significant antimicrobial-resistant pathogens both in the United States and poses severe challenges to public health and healthcare systems. While methicillin-susceptible *Staphylococcus aureus* (MSSA) has historically been a leading cause of bloodstream-associated infections (BAIs), MRSA has emerged as a greater threat due to its antibiotic resistance and clinical impact (pneumonia, septic thrombophlebitis, necrotizing fasciitis) that has led to over 70,000 infections and 9,000 associated deaths in 2025 [2, 4]. MRSA infections are also linked with higher mortality rates, increased hospital stays, and economic burden (2-3 times more expensive than MSSA infections). As a result, MRSA BAIs are not only threats in the clinical space, but also affect healthcare systems economically and operationally [9].

How Antibiotic Resistance Works. The rapid emergence of resistance with MRSA appeared just two years after methicillin's introduction and highlights how quickly *S. aureus* can evolve under antibiotic pressure, making it essential to understand how resistance develops[5]. Mlynarczyk-Bonikowska et al. describes how MRSA becomes resistant to antibiotics in several ways. First, it can borrow resistance genes from other bacteria, which gives it protection against entire classes of drugs like penicillins or vancomycin. Secondly, it can gradually adapt over time, developing small changes that make antibiotics less effective. Finally, MRSA also has built-in survival tactics like pumps that remove drugs out of its cells, produce enzymes that break them down, or create biofilms to hide behind so antibiotics struggle to reach it [8]. These strategies stack together, making MRSA extremely hard to eliminate. While reviews describe these mechanisms well, they rarely explain which ones matter most in real-world infections, which is why studies linking genetics to patient outcomes are essential.

Shifting Epidemiology of MRSA Lineages. The traditional distinction between hospital-acquired MRSA (HA-MRSA) and community-acquired MRSA (CA-MRSA) is becoming unclear. CA-MRSA was once considered more treatable, carrying fewer resistance genes and often producing toxins like PVL that primarily caused skin infections [10]. In contrast, HA-MRSA was defined by extensive drug resistance and its association with invasive disease. However, recent studies show increasing overlap between the two groups. CA-MRSA strains are appearing in hospitals, while HA-MRSA strains persist in community settings. This convergence makes treatment more challenging, as physicians can no longer assume resistance patterns based on where an infection originated [10]. Ongoing genomic surveillance is therefore needed to track resistance trends and guide therapy. While biological convergence between MRSA lineages complicates treatment decisions, infection risk is further shaped by socioeconomic and environmental conditions.

Impact of Social Determinants of Health on MRSA. While comorbidities play a significant role in MRSA infections, recent studies have linked broader social and environmental factors to MRSA susceptibility and outcomes [2, 6]. Review of the existing literature highlights how social determinants of health (SDOH) shapes the epidemiology and outcomes of MRSA infections [2, 9]. Risk factors include living in crowded or unsanitary conditions, a history of incarceration, injection drug use, smoking, and previous trauma, as well as chronic health conditions that are more common in socioeconomically disadvantaged populations, such as congestive heart failure, HIV, and obesity [2]. Furthermore, CA-MRSA is seen to spread more easily in densely populated urban areas with higher rates of homelessness, incarceration, and household crowding [9]. Analysis at the neighborhood level also reveals that individuals with community-acquired MRSA Skin and Soft Tissue Infections (SSTIs) are more likely to have household incomes that are $20,000 less than those without such infections [2], and analysis at the demographic level shows that African Americans face twice the risk of infection compared to other racial and ethnic groups, even as overall MRSA rates decline [6]. A further analysis of the literature

[*] All authors contributed equally to this research.

highlights the need for deeper analysis to determine if a causal link exists between SDOH and MRSA.

Gaps in Analysis of Social Determinants. Although substantial literature links socioeconomic factors to MRSA risk, most existing studies are associational and do not establish the causal mechanisms that connect these factors to MRSA outcomes [2, 6]. This makes it challenging to develop effective strategies that address the unique vulnerabilities of specific subpopulations, including older adults and higher-risk demographic and socio-economic groups [6]. Additionally, analysis at different geographic levels (counties, census tracts, ZIP codes) make it difficult for comparison across results [2, 9]. Methodological challenges also exist, including difficulties distinguishing MRSA from MSSA infections due to International Classification of Diseases (ICD) coding in datasets, and potential geographic biases that may influence spatial analysis[9].

Modeling Approaches. Previous research has evaluated the association between California neighborhood-level SSTI risk and population, neighborhood poverty rates, and healthcare shortage areas through a Bayesian Poisson regression model [9]. Time series analyses have also been performed to gain insight into decadelong seasonal patterns and pandemic-associated changes in hospital onset MSRA infections at a tertiary care referral center [7]. Other literature has employed multivariate analysis to investigate the association between MRSA bloodstream infections as well as socioeconomic factors and antibiotic prescriptions at the county level. This research found both of these to be statistically significant in influencing infection rates [1]. With the data used for this project and research problem, each of these techniques makes sense to exploit as this project explore how differences in social and structural resources across localities influenced MRSA infection patterns and treatment outcomes in California over the past decade, and to what extent these factors can be used to predict future infection risk at the county level. Poisson regression would be useful to model arrivals of MRSA infections, time series would allow us to study infections year over year, and multivariate analysis is sensible given the variety of factors in the project dataset (patient days, county, facility type, etc.). However, each of these approaches comes with its own set of weaknesses as well. Poisson and multivariate regression models would not capture the temporal effect that is potentially present in influencing the MRSA infection rate. Additionally, time series analysis on its own would be too simplistic to help understand the effects socioeconomic factors have on Californian county MSRA infection trends.

2 Problem Statement

2.1 Layman Problem Definition

Differences in local socioeconomic conditions (income inequality, unemployment, poverty level, access to healthcare providers) contribute to why MRSA infection trends vary across regions. Focusing specifically on California, the goal of this project is to study whether these factors are related to infection patterns and predict future infections. This could help public health agencies target resources and interventions more effectively.

2.2 Formal Problem Definition

This project aim to investigate whether variations in local socioeconomic and healthcare resource factors over the past decade can explain differences in MRSA infection trends across California counties. Formally, let y_{it} represent the number of MRSA bloodstream infections in county i at time t, with:

$$y_{it} = f(X_{it}, S_i, H_i) + \epsilon_{it}$$

where X_{it} are time-varying socioeconomic features (e.g., poverty rate, population density), S_i are static structural factors (e.g., hospital availability), and H_i are healthcare resource variables. The objectives are to (1) quantify the relationship between these factors and MRSA infection trends, and (2) predict future infection risk at the county level.

3 Proposed Methodology

This project will employ a combination of statistical and machine learning approaches to model MRSA infection risk and identify key predictors:

- **Spatiotemporal Deep Learning and Regression Models:** By treating counties as nodes and borders between counties as edges, graph convolution can be performed to mix neighbor signals and then feed the sequence into a recurrent neural network.
- **Time-Series Forecasting:** Rolling-horizon approaches, such as ARIMA or SARIMA, can be used to analyze temporal trends.
- **Machine Learning Models:** Random forest and gradient boosting will be important in capturing nonlinear relationships and ranking the importance of predictors.

This combination allows us to (1) model infection incidence, (2) understand yearly and seasonal trends, and (3) evaluate the relative importance of socioeconomic and healthcare variables.

Each modeling approach contributes distinct analytical strengths to the project modeling framework. The spatiotemporal deep learning model leverages graph convolution to capture inter-county dependencies—such as patient movement or shared healthcare infrastructure, while the recurrent layer (e.g., LSTM or GRU) learns temporal evolution patterns in infection rates. This allows the model to dynamically propagate information from both spatial and temporal contexts. The time series models serve as interpretable baselines for forecasting, capturing linear temporal dependencies and seasonal fluctuations that can help validate or contextualize deep learning outputs. Finally, tree-based machine learning models like random forest and gradient boosting provide complementary flexibility in handling complex nonlinear relationships and heterogeneous predictors, such as socioeconomic variables, hospital capacity, and antibiotic prescribing rates. Together, these techniques balance interpretability, predictive performance, and temporal-spatial awareness within a unified analytical pipeline.

3.1 Risks and Payoffs, Costs, and Length of Project

The risks associated with this research project are potential runtime issues and lack of result interpretability. One payoff, however, is the creation of easy-to-understand models that clearly indicate how differences in California socioeconomic conditions help explain why MRSA infection trends vary from county to county. Additionally, this project hopes to develop findings that can help public health agencies in California target resources an interventions for MRSA treatment more effectively. The financial costs associated with this project are likely minimal to none. Computational (runtime) costs, though expected to be low, are the only potential costs identified for project work. This project will be completed over the remainder of the Fall 2025 semester - in total, about two months.

4 Evaluation Strategy

This project will employ a comprehensive suite of evaluation techniques to assess model performance, interpretability, and robustness across spatial and temporal dimensions.

1. **Predictive Performance Metrics:** Since the outcome feature of interest, the MRSA infection rate, is continuous, accuracy of models will be assessed using the root mean squared error (RMSE), the mean absolute error (MAE) and the coefficient of determination (R^2). RMSE will penalize large deviations more strongly, emphasizing high-error regions, while MAE provides a more robust measure against outliers. R^2 will quantify the proportion of variance explained by each model, allowing for direct compare performance across methods such as graph-based deep learning, ARIMA/SARIMA, and tree-based regression models.
2. **Temporal and Spatial Validation:** Given the spatiotemporal nature of MRSA incidence, rolling-origin cross-validation will be adapted to evaluate temporal robustness. This involves training on data up to time t and testing on time t+1, mimicking real-world forecasting. To assess spatial generalization, leave-one-region-out cross-validation will be utilized, where entire counties or groups of counties are excluded from model training. This approach tests whether developed models can accurately predict infection patterns in previously unseen geographic areas—a critical step for regional public health planning.
3. **Model Interpretability and Feature Importance:** Understanding the drivers of infection is as important as prediction accuracy. For tree-based models such as random forests and gradient boosting, permutation importance and SHAP (SHapley Additive exPlanations) will be computed values to rank predictors on their explainability and uncover nonlinear or interaction effects among socioeconomic and healthcare-related variables. For spatiotemporal deep learning models, node embeddings will be analyzed and attention or saliency maps to interpret how temporal sequences and spatial neighborhoods influence infection dynamics.
4. **Uncertainty and Calibration Analysis:** To quantify confidence in predictions, calculate prediction intervals (e.g., 95% coverage) will be calculated and bootstrapping will be performed to estimate uncertainty around model parameters and predicted infection rates. For time-series models like ARIMA/SARIMA, residual diagnostics and autocorrelation functions (ACF/PACF) will also be examined to ensure that temporal dependencies are appropriately captured. Well-calibrated uncertainty estimates will improve the reliability of forecasts used for resource allocation or policy evaluation.

4.1 Midterm and Final "Exams" to Check for Success

In addition to quantitative model evaluation, two key project checkpoints are established to assess progress and success. The midterm evaluation will focus on verifying data readiness, the functionality of the modeling pipeline, and the ability of baseline models (e.g., ARIMA/SARIMA, Random Forest) to capture preliminary spatial and temporal patterns in MRSA incidence. Success at this stage will be measured by improved predictive accuracy over naive benchmarks and evidence that the models learn meaningful regional trends. The final evaluation will emphasize model refinement, interpretability, and policy relevance—ensuring that tuned models generalize across both time and geography, provide well-calibrated uncertainty estimates, and yield actionable insights into the social and healthcare determinants of MRSA risk. Collectively, these checkpoints will serve as practical “exams” to validate that the project’s technical outputs align with its broader epidemiological and public health objectives.

5 Data Description

MRSA Infection Data

- **Source:** California Department of Public Health [3]
- **Format:** CSV, annual hospital-reported infection counts and rates over approximately 12 years
- **Key Features:** Hospital identifiers, infection counts, county, reporting period

Socioeconomic and Health Determinants Data

- **Source:** County Health Rankings [11]
- **Format:** database with yearly CSV’s, county-level indicators such as poverty, education, housing, and healthcare access

6 Expected Outcomes

- A fully cleaned and merged dataset combining MRSA infection data with socioeconomic and healthcare factors.
- Exploratory spatial and temporal analyses, including visualizations and trend analyses.
- One or more predictive models (e.g., spatial regression, time series, machine learning) with evaluated performance.
- Comprehensive model evaluation using RMSE, MAE, and R^2 metrics, along with spatial and temporal cross-validation to assess generalizability.
- Interpretability analyses using SHAP values, permutation importance, and saliency maps to identify key predictors and understand model behavior.
- Quantified uncertainty and calibration assessments (e.g., prediction intervals, residual diagnostics) to ensure robust and trustworthy predictions.
- A ranked list of the most influential social determinants of health (SDOH) factors associated with MRSA risk.
- County-level infection risk forecasts and/or risk maps.

- A written report detailing methods, results, model comparisons, limitations, and public health implications.

7 Team Responsibilities and Timelines

- **Project Proposal**
  - Hannah: Reviewed literature on MRSA infection trends and epidemiology, formatted and structured the LaTeX report.
  - AJ: Reviewed literature on different modeling approaches and wrote the methodology and evaluation sections.
  - Alan: Reviewed literature on social determinants of health (SDOH) and their impact on MRSA, and sourced and compiled datasets.
- **Project Milestone Report**
  - Hannah: Develop visualizations and performed trend analysis, help with data exploration, and draft the report.
  - AJ: Build initial models, help prepare the data, and contribute to interpreting results.
  - Alan: Perform exploratory data analysis, help analyze trends, and work on early modeling.
- **Final Report**
  - Hannah: Write results and discussion sections and connect findings to public health insights.
  - AJ: Improve and test models, summarize limitations, and help combine findings from data and public health perspectives.
  - Alan: Finalize model results, analyze key variables, and interpret trends.

References

[1] Nikolaos Andreatos, Fadi Shehadeh, Elina Eleftheria Pliakos, and Eleftherios Mylonakis. 2018. The impact of antibiotic prescription rates on the incidence of MRSA bloodstream infections: a county-level, US-wide analysis. *International journal of antimicrobial agents* 52, 2 (2018), 195–200.

[2] Sarah Blackmon, Esther E Avendano, Sweta Balaji, Samson Alemu Argaw, Rebecca A Morin, Nanguneri Nirmala, Shira Doron, and Maya L Nadimpalli. 2025. Neighborhood-level income and MRSA infection risk in the USA: systematic review and meta-analysis. *BMC Public Health* 25, 1 (2025), 1074.

[3] California Department of Public Health, Healthcare-Associated Infections Program. 2025. Methicillin-resistant Staphylococcus aureus (MRSA) bloodstream infections (BSI) in California hospitals. https://data.chhs.ca.gov/dataset/methicillin-resistant-staphylococcus-aureus-mrsa-bloodstream-infections-bsi-in-california-hospitals.

[4] Centers for Disease Control and Prevention. 2025. Infection control guidance: Preventing methicillin-resistant *Staphylococcus aureus* (MRSA) in healthcare facilities. https://www.cdc.gov/mrsa/hcp/infection-control/index.html. U.S. Department of Health and Human Services.

[5] Mark C Enright, D Ashley Robinson, Gaynor Randle, Edward J Feil, Hajo Grundmann, and Brian G Spratt. 2002. The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). *Proceedings of the National Academy of Sciences* 99, 11 (2002), 7687–7692.

[6] Inyoung Jun, Sarah E Ser, Scott A Cohen, Jie Xu, Robert J Lucero, Jiang Bian, and Mattia Prosperi. 2023. Quantifying health outcome disparity in invasive methicillin-resistant staphylococcus aureus infection using fairness algorithms on real-world data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*. World Scientific, 419–432.

[7] Pedro Martínez-Ayala, Judith Carolina De Arcos-Jiménez, Adolfo Gómez-Quiroz, Brenda Berenice Avila-Cardenas, Roberto Miguel Damian-Negrete, Ana María López-Yáñez, Leonardo García-Miranda, Carlos Roberto Álvarez-Alba, and Jaime Briseno-Ramirez. 2025. Time-Series Analysis of Staphylococcus aureus and MRSA Trends, Seasonality, and Pandemic-Associated Disruptions in a Tertiary-Care University Hospital (2016–2025). (2025).

[8] Beata Mlynarczyk-Bonikowska, Cezary Kowalewski, Aneta Krolak-Ulinska, and Wojciech Marusza. 2022. Molecular mechanisms of drug resistance in Staphylococcus aureus. *International journal of molecular sciences* 23, 15 (2022), 8088.

[9] Brittany L Morgan Bustamante, Laura Fejerman, Larissa May, and Beatriz Martínez-López. 2024. Community-acquired Staphylococcus aureus skin and soft tissue infection risk assessment using hotspot analysis and risk maps: the case of California emergency departments. *BMC Public Health* 24, 1 (2024), 123.

[10] Haiying Peng, Dengtao Liu, Yuhua Ma, and Wei Gao. 2018. Comparison of community-and healthcare-associated methicillin-resistant Staphylococcus aureus isolates at a Chinese tertiary hospital, 2012–2017. *Scientific reports* 8, 1 (2018), 17916.

[11] University of Wisconsin Population Health Institute. 2025. County Health Rankings: California Data and Resources. https://www.countyhealthrankings.org/health-data/california/data-and-resources.