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

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

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

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

Abstract

Methicillin-resistant *Staphylococcus aureus* (MRSA) remains a major antimicrobial resistance threat in the United States, yet the influence of social and structural conditions on infection patterns is not well understood. This study examines how county-level social determinants of health (SDOH) and healthcare exposure relate to geographic and temporal variation in MRSA bloodstream infections across California from 2017 to 2024. We integrate hospital reported MRSA surveillance data with SDOH indicators and patient-day exposure metrics, and evaluate statistical, time series, and machine learning models to predict county-level infection burden. A Random Forest model incorporating lagged infection counts and SDOH features achieved the strongest performance (R^2 = 0.902, RMSE = 9.0), outperforming linear, ARIMA, boosting, and deep learning approaches. Feature importance analysis showed that population size, healthcare utilization, particularly patient-days in major teaching hospitals and prior year infection burden were the dominant predictors, while most SDOH variables contributed limited independent signal due to high multicollinearity and temporal instability. These findings indicate that MRSA risk is primarily driven by population scale, healthcare exposure, and persistent year-to-year transmission patterns, with structural SDOH playing a secondary role. The modeling framework provides a basis for targeted surveillance and forecasting efforts aimed at counties and facility types with the highest sustained MRSA burden.

1 Introduction

Invasive infections caused by methicillin-resistant *Staphylococcus aureus* (MRSA) impose a persistent burden on U.S. healthcare systems. While the biological mechanisms of MRSA are well documented, less is known about how community-level social and structural conditions shape infection patterns. In particular, the influence of social determinants of health (SDOH) such as income, housing stability, healthcare access, and population density on geographic and temporal trends in MRSA infections remains underexplored. This study examines how socioeconomic factors contribute to variations in MRSA bloodstream infection rates across California counties over the past decade. We integrate hospital-reported infection data with county-level demographic and healthcare indicators, applying regression, time series, and machine learning models to assess infection risk. Our analysis identifies key predictors, generates county-level forecasts, and visualizes spatial and temporal trends to inform targeted public health interventions. By linking structural conditions to MRSA epidemiology, this work addresses critical analytical gaps and supports more equitable, data driven infection control strategies.

2 Related Work

2.1 Clinical and Biological Context

Background and Significance. Since MRSA infections are so prevalent, it poses severe challenges to public health and healthcare systems. While methicillin-susceptible S. aureus (MSSA) has historically been a leading cause of bloodstream-associated infections (BAIs), MRSA has emerged as a greater threat due to its resistance and clinical impact, causing over 70,000 infections and 9,000 deaths in 2025 [2, 5]. MRSA infections are associated with higher mortality, longer hospital stays and substantially higher costs, often 2-3 times that of MSSA infections [12]. These outcomes underscore MRSA's dual burden on patients and healthcare systems.

How Antibiotic Resistance Works. MRSA resistance emerged just two years after methicillin's introduction, showing how rapidly S. aureus evolves under antibiotic pressure [8]. MRSA gains resistance by borrowing genes from other bacteria, gradually adapting through small mutations and using intrinsic mechanisms like efflux pumps, enzyme degradation and biofilm formation [11]. Many of these traits are carried on mobile genetic elements such as plasmids and transposons, allowing resistance to spread quickly across strains and even species. The biofilm phenotype further enhances survival by reducing antibiotic penetration and protecting dormant cells that can repopulate once treatment ends. Together, these strategies make MRSA difficult to eliminate.

Shifting Epidemiology of MRSA Lineages. The traditional distinction between hospital-acquired (HA-MRSA) and community-acquired (CA-MRSA) strains is blurring. CA-MRSA, once more treatable, now appears in hospitals, while HA-MRSA persists in communities. CA-MRSA is typically linked with toxins such as Panton-Valentine leukocidin (PVL), whereas HA-MRSA carries broader resistance genes. Increasing genetic exchange between lineages has produced hybrid strains that circulate across both settings. This overlap complicates treatment since clinicians can no longer rely on the origin of the infection to predict resistance [14]. Ongoing genomic surveillance is needed to track these trends. Patient movement between hospitals and communities further accelerates this mixing, highlighting the need to consider the social and environmental context when assessing infection risk.

2.2 Social and Environmental Influences on MRSA

Impact of Social Determinants of Health on MRSA. Beyond individual comorbidities, broader social and environmental factors strongly influence MRSA susceptibility and outcomes [2, 9]. Social determinants of health (SDOH) such as crowding, incarceration, injection drug use and limited healthcare access shape the epidemiology of infection [2, 12]. Chronic conditions more common in socioeconomically disadvantaged populations such as heart failure, HIV and obesity further elevate risk. CA-MRSA also spreads more readily in densely populated areas with higher rates of homelessness and household crowding [12]. Individuals with community-acquired skin and soft tissue infections (SSTIs) often have household incomes roughly $20,000 lower than those without such infections [2] and African Americans face nearly twice the infection risk of other racial and ethnic groups, even as overall MRSA rates decline [9]. These disparities highlight the need to clarify whether social determinants exert a causal influence on MRSA burden.

3 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:

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

where X_{it} are time-varying socioeconomic features (e.g., poverty rate, population density), S_i are static structural factors (e.g., hospital patient days) . 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.

4 Proposed Method

4.1 Why our Method Improves on Previous Approaches

Although previous studies have linked socioeconomic factors, infection counts and temporal patterns to MRSA trends, most methods remain limited to a single analytical dimension. Prior work has relied on cross-sectional associations between SDOH and MRSA outcomes [2, 9], but these models do not jointly incorporate spatial, temporal and facility-level variation, making it difficult to capture the mechanisms that connect local conditions to yearly fluctuations in infection burden. In addition, differences in geographic resolution and ICD-based misclassification can introduce bias that reduces predictive accuracy [12].

Similarly, while Bayesian Poisson regression has been effective for modeling neighborhood SSTI incidence [12], time-series analyses have characterized seasonal and pandemic-related trends in hospital-onset MRSA [10], and multivariate analyses have identified relationships between bloodstream infections, socioeconomic factors and antibiotic use [1], each approach captures only one aspect of MRSA epidemiology. None of these methods fully address the nonlinear interactions, multicollinearity and lagged temporal structure that arise when modeling county-level MRSA using both SDOH and facility-level hospital data.

Our modeling framework directly addresses these methodological gaps by combining spatial, temporal and socioeconomic information into a single modeling approach. Random Forests are especially effective in small-n, large-p settings like ours and handle the heavy multicollinearity present in SDOH variables—consistent with our PCA results showing that most variation collapses into only a few components. Incorporating lag features further helps stabilize noisy year-to-year survey measurements and adds temporal structure that earlier SDOH focused studies did not include. In addition, tree-based models naturally learn hospital specific interactions through their splitting rules without requiring us to manually engineer interaction terms. We also evaluate the models using rolling-origin cross-validation which reflects realistic forward-in-time forecasting rather than static cross-sectional comparison. Finally, we restrict hyperparameter tuning to the best-performing model configuration, reducing the risk of overfitting the entire pipeline.

Taken together, this unified approach fills the analytical gaps in previous work by capturing nonlinear relationships, lagged temporal effects and facility-level heterogeneity, leading to more accurate and interpretable predictions of county-level MRSA infection risk.

4.2 Data Prep & Feature Engineering

4.2.1 Datasets. We assembled a comprehensive county-level dataset spanning 2017–2024 by merging two primary sources. The first source is a MRSA (Methicillin-resistant *Staphylococcus aureus*) infection surveillance dataset, which provides annual counts of MRSA infections for each county [4]. The second source is a collection of county-level Social Determinants of Health (SDOH) indicators over the same time frame [16]. The SDOH data includes a broad array of socio-economic, demographic, and environmental variables (for example, median income, housing characteristics, educational attainment, population demographics, and air quality measures). Combining these datasets on a common county-year basis allowed us to study how MRSA infection trends relate to underlying community factors.

4.2.2 Target Variables. We defined two sets of target outcomes for prediction. The primary target is the total number of reported MRSA infections in each county per year. In addition, we examined facility-specific infection counts stratified by hospital type within each county-year. This means that for each category of healthcare facility (e.g. different hospital types or settings) we tracked the number of MRSA infections occurring in that category. Modeling both the overall infection count and the disaggregated facility specific counts provides a more detailed understanding of MRSA incidence patterns across different healthcare environments.

4.2.3 Input Features. Our modeling incorporated a rich set of input features capturing temporal context, geographic location, healthcare exposure, community socioeconomic factors, and historical infection levels. The temporal feature is simply the year of observation (2017 through 2024) which can capture overall time trends or shocks affecting all counties. The geographic features consist of one hot encoded indicators for each state and each county. By including these, the model can account for region specific effects and unobserved heterogeneity (such as differences in healthcare infrastructure or reporting practices) at the state or county level. As a proxy for healthcare utilization or exposure opportunity, we included patient-days for each hospital type in the county (the total number of patient-days within each type of healthcare facility). This feature reflects the volume of patients and healthcare interactions in that county and is expected to correlate with infection risk (more patient-days could lead to more opportunities for MRSA transmission). We also incorporated a comprehensive set of SDOH indicators as continuous features. These include economic measures (like median household income and employment rates), housing and living conditions (such as housing density or overcrowding rates), population demographics (e.g. age distribution, racial/ethnic composition), educational attainment levels, healthcare access metrics (like the percentage of uninsured individuals), and environmental quality indicators (for example, annual average PM2.5 concentration as a measure of air pollution). Finally, we introduced lagged infection features: for each county and each facility type, we included the prior year's MRSA infection count as an input. These lagged features capture temporal dependencies under the assumption that MRSA incidence in a given year may be partially predicted by the burden in the previous year (due to factors like persistent local infection sources or slow changing community risk factors).

4.2.4 Exploratory Data Analysis (EDA) Insights. Before building predictive models, we conducted exploratory analyses to understand the data's characteristics. MRSA Incidence Trends: We found that MRSA infection rates (expressed per 1,000 patient-days) were relatively low but non-negligible, generally ranging between approximately 0.04 and 0.06. There was a notable temporal pattern in the data: a sharp dip in MRSA infection incidence occurred in 2020, followed by a rebound in 2021 and subsequent years. In other words, many counties saw lower MRSA rates during 2020, then an increase back toward pre-2020 levels after that year. This trend aligns with the timeline of the COVID-19 pandemic, during which heightened infection control practices and changes in healthcare utilization may have transiently suppressed MRSA transmission; by 2021, as healthcare operations normalized, MRSA rates began to rise again.

4.2.5 SDOH Feature Structure. We used principal component analysis (PCA) to investigate the structure of the SDOH feature space [13]. The PCA results indicated that the numerous SDOH variables could be effectively summarized by three major components. These top components appeared to correspond to (1) socioeconomic status, capturing variation in income, education, and related economic indicators; (2) housing and demographic composition, reflecting aspects like housing stability, crowding, and population makeup; and (3) urban density, which separated urbanized counties from rural ones (with loadings from variables related to population density and possibly infrastructure). This analysis confirmed that many of the SDOH features were correlated with one another and tended to cluster around underlying common themes. In fact, we observed strong multicollinearity among the raw SDOH features: for instance, counties with high median income often also had high education levels and lower uninsured rates, while densely populated counties tended to have similar housing and pollution profiles. Such interrelationships informed our feature engineering and modeling strategy, since models can be prone to instability or overfitting when many collinear predictors are included. We accounted for this either by reducing dimensionality (using approaches like PCA or feature selection) or by choosing models robust to multicollinearity.

4.2.6 Temporal Variability. Our EDA also highlighted that some features exhibited considerable year-to-year variability at the county level. In particular, variables like annual PM2.5 (air pollution level) and the percentage of uninsured individuals were not static, they fluctuated over time and sometimes substantially. This instability means that using the raw year specific values could introduce noise. A sudden one year spike or drop might mislead a model into overemphasizing a transient change. To address this, we considered strategies such as smoothing these time series or using lagged values. For example, instead of using the current year's PM 2.5 value alone, one might include a multi year average or the previous year's value to dampen short term fluctuations. Similarly, the inclusion of lagged infection counts helps the model account for temporal dependencies and possibly filter out irregular year to year jumps in incidence by anchoring predictions to the recent historical baseline.

4.2.7 Pre-processing Steps. We constructed three main analysis datasets, each organized at the county–year level from 2017 - 2024:

- **(1) Base MRSA panel (Base / No SDOH)**
  This dataset contains one row per county and year, including:
  - total MRSA bloodstream infections
  - facility-specific infection counts for each hospital type
  - patient-day totals for each hospital type
  - geographic identifiers (State, County, Year)
- **(2) SDOH-augmented panel (SDOH)**
  This data set was created by merging the base MRSA panel with the social determinants of health dataset using a left join on State, County, and Year. This adds time-varying socioeconomic attributes such as income, education, housing, and insurance coverage.
- **(3) Lag-augmented panel (SDOH + Lag)**
  This dataset was built from the SDOH-augmented panel by sorting the data and constructing one-year lagged MRSA infection variables. For each hospital type, the prior-year infection count was assigned to the current year using a grouped one-year shift within each county. County-year rows without a prior-year value (the first year observed for each county) were removed. This produces a dataset with complete lagged features for use in models that incorporate facility and county prior-year MRSA burdens.

4.3 Modeling Approach

We evaluated a range of statistical and machine learning models to determine which approach most effectively predicts county-level MRSA infections. Each model was trained to forecast both total infections and facility-specific infections. The configurations were:

(1) **Linear Regression (lr) – Base, SDOH:** Ordinary least-squares regression models used as simple baselines. One model included only Year, State, County, and patient-day variables; another incorporated SDOH variables. These models help assess linear separability in the feature space [13].
(2) **ARIMA (arima) – Base:** A univariate time-series baseline applied separately to each county, using infection counts across years. ARIMA captures temporal patterns but does not exploit cross-county structure or high-dimensional predictors [15].
(3) **Random Forest (rf) – Base:** Baseline tree ensemble using only temporal, geographic, and patient-day exposure features. This configuration provides a capacity-controlled nonlinear baseline [3].

(4) **Random Forest (rf) – SDOH:** Extends the baseline RF by incorporating Social Determinants of Health (SDOH) indicators (income, demographics, education, air quality, housing). This model evaluates whether community-level risk factors improve predictive accuracy [3].
(5) **Random Forest (rf) – SDOH + Lag:** Adds lagged infection features to the SDOH model. A lag-1 infection variable represents the previous year's infection count for the same county and facility type, allowing the model to capture temporal persistence and autoregressive structure in MRSA incidence [3].
(6) **XGBoost Regressor (xgb) – SDOH + Lag:** Gradient-boosted trees trained on the same full feature set. Boosting provides an alternative nonlinear learner to compare against Random Forest performance [3, 6].
(7) **Deep Learning (keras) – SDOH + Lag:** A feed-forward neural network with two hidden layers (128 and 64 units, ReLU activation, dropout regularization). This model tests whether high-capacity function approximators outperform tree-based ensembles under the same feature conditions [7]

4.4 Training Strategy

Two modeling tasks were performed for every configuration:

(1) Total infections model: Predicts the total county-level MRSA infections for 2024.
(2) Facility-specific models: A separate model for each hospital type (Critical Access, Major Teaching, etc.), allowing fine-grained evaluation across healthcare settings.

Categorical features (State, County) were one-hot encoded. Numeric features (patient-days, SDOH, lagged infections) were passed through without scaling. Lagged models dropped the first year for each county since lag-1 values were unavailable.

4.5 Overall Model Comparison

To determine which model class and feature configuration worked best, all models were evaluated on the 2024 test set. Several clear patterns, shown in Table 1, emerged across experiments, which ultimately guided the selection of the final model.

Model	MAE	R^2	RMSE
lr_no_sdoh	4.9035	0.9077	9.9405
lr_with_sdoh	183.4025	-122.8151	364.0872
arima_no_sdoh	2.4042	0.9815	4.4551
rf_no_sdoh	3.7461	0.8986	10.4178
rf_with_sdoh	6.9664	0.8631	12.1065
rf_with_sdoh_lag	6.2107	0.9026	10.2124
xgb_with_sdoh_lag	5.8166	0.8461	12.8367
keras_with_sdoh_lag	12.7254	-0.1505	35.0967

Table 1: Performance metrics across model configurations for the 2024 test set.

4.5.1 Evaluation Metrics. All evaluation metrics were computed on count data, since the outcome represents the number of MRSA bloodstream infections in each county–year. To assess model performance, three metrics were used to capture a different property of predictive accuracy and behaves differently when applied to discrete, low-count outcomes.

- MAE measures the average absolute difference between predicted and observed infection counts. Because many counties report only a small number of cases, MAE is an especially relevant measure; an error of one or two infections can represent a large proportion of the true count. MAE therefore provides an interpretable, scale-consistent evaluation of accuracy for this type of sparse count data.
- RMSE also reflects the magnitude of prediction errors but penalizes larger mistakes more heavily due to the squared term. This makes RMSE useful for identifying models that occasionally produce large deviations, even if their average error is relatively small.
- R^2 quantifies how much of the variation in infection counts the model explains relative to a baseline model predicting the mean. While R^2 is informative for higher-count or more variable outcomes, it becomes unstable in very low-count settings. In counties with only 0–2 observed infections, small changes in prediction can cause R^2 to appear near zero or even negative, despite the model having low MAE and RMSE. For this reason, R^2 is interpreted cautiously for rare-event settings, and greater emphasis is placed on MAE and RMSE in those cases.

4.5.2 General Performance Trends. Models that excluded SDOH features often performed better than those that incorporated them. This appears to be due to the structure of the SDOH dataset: many of the socioeconomic measures were highly collinear and therefore did not introduce meaningful independent signal for the models to learn from. This was especially visible in the linear models, where multicollinearity directly degrades coefficient stability and predictive performance upon training on the SDOH dataset. In contrast, model accuracy improved when lagged infection features were added. The lag variables capture the previous year's MRSA burden, which is a strong and pseudo-temporal predictor of current infection levels.

Across all classes of models, the tree-based approaches (specifically XGBoost and Random Forest) were the most robust performers. These models naturally handle nonlinear relationships, correlated predictors, and mixed feature types, making them better suited to the structure of this dataset than the linear or deep learning models. The neural network configuration used in this study performed the worst overall, likely due to the relatively small number of county-year observations and the highly correlated nature of the features, which make deep models prone to overfitting.

Although ARIMA performed well for the aggregate series, it models only a single time series and cannot represent cross-county differences or spatial patterns. The SDOH variables we passed into the model also added little signal due to their high collinearity and limited variation over time. For these reasons, ARIMA was not suitable for forecasting the full county-level panel.

4.5.3 Model-by-Model Performance Summary. Our Linear Regression model performed best without the SDOH data added, with an $R^2 \approx 0.91$. When adding the SDOH data the R^2 value tanked to -123 because of the extreme multicolinearity as discussed above. We excluded this model because the relationship between predictors and infections is clearly nonlinear and linear regression is too fragile for our feature correlations.

Although ARIMA produced excellent looking forecasts for total infections when evaluated one county at a time ($R^2 \approx 0.98$), these results do not generalize to our full prediction task. ARIMA can only use a single univariate series at a time, meaning we must train a separate model for every county and the model cannot access patient-day counts, SDOH variables, or facility-specific information. Because ARIMA only learns from the past values of its own series, it mainly captures each county's historical trend rather than learning shared patterns across counties. This makes the model highly prone to overfitting and prevents it from modeling the spatial, demographic, and healthcare-utilization factors that drive MRSA variation. For these reasons, ARIMA was not suitable for the final modeling framework.

Our XGBoost had moderate overall performance with an $R^2 \approx 0.85$. These boosting methods usually excel with large datasets and with our smaller dataset, they have a harder time learning. XGBoost is also more sensitive to noise in year to year fluctuations than Random forest models, which is something that happens a lot in our dataset. We ultimately decided against using this model in favor of Random Forest, which we found has a higher accuracy, smoother predictions and more reliable generalization.

Our neural network model exhibited extremely poor test performance ($R^2 \approx -0.15$), indicating that it failed to learn stable or generalizable patterns from the available data. Neural networks generally need large amounts of data, but in our case each county-year is only one observation, giving the model far too few examples to learn from. The feature space is also highly correlated, which makes training even harder. Because tree-based models performed much better with the same inputs, we excluded the neural network from further analysis. The model was simply too complex for the size and structure of our dataset, leading to unstable and unreliable predictions.

4.6 Best Performing Model

Random Forest with SDOH + Lag. The best-performing model in our study was the Random Forest trained on both SDOH variables and lagged infection counts. On the 2024 test set, this model achieved strong accuracy ($R^2 \approx 0.902$, $MAE \approx 5.4$, $RMSE \approx 9.0$). The inclusion of lagged infections was the key factor behind its superior performance: a lag-1 feature provides each county's infection burden from the previous year, allowing the model to capture the temporal persistence characteristic of MRSA trends—something SDOH variables alone cannot provide.

4.6.1 Hyperparameter Tuning and Validation. Hyperparameter tuning further improved model stability, yielding a final **Random Forest with SDOH + Lag (TUNED) model** that used 400 trees, maximum depth = 10, minimum leaf size = 4, and 80% feature subsampling. This tuned configuration showed modest improvements in error and R^2 compared to the untuned model. To evaluate temporal robustness, we applied rolling-origin cross-validation—retraining the model on progressively more years and predicting the next. Across these rolling forecasts, performance increased steadily as more historical data became available: R^2 improved from approximately 0.72 for 2020 predictions to roughly 0.924 for 2023–2024, even through pandemic related variability. Together, these results confirm that combining SDOH context with lagged infection history provides the Random Forest with both the cross-sectional and temporal structure necessary for reliable MRSA forecasting.

5 Experiments and Results

This section summarizes the key epidemiologic and modeling findings based on the tuned Random Forest model with SDOH and lagged infections, unless otherwise noted. The results are organized to answer the core scientific questions motivating the study: how MRSA burden is distributed across California, which hospital types show predictable patterns, which demographic and structural factors drive infection risk, and how model performance reflects underlying epidemiology. Each subsection integrates statistical outcomes with epidemiologic interpretation.

5.1 Questions to Answer

- How are MRSA infections distributed across California counties, and which areas account for the greatest burden (Section 5.2.2)?
- Which hospital types exhibit predictable vs. unpredictable MRSA patterns, and what drives these differences (Section 5.2.1)?
- Which demographic, structural, and healthcare-utilization features most strongly influence MRSA incidence across counties (Section 5.2.1, 5.2.3)?
- Do Social Determinants of Health (SDOH) meaningfully contribute to MRSA prediction after accounting for population size and healthcare exposure (Section 5.2.4)?
- Which county-level demographic or structural factors (SDOH) are most associated with elevated risk (Section 5.2.3)?
- How do linear, time-series, ensemble, and neural models compare in their ability to capture MRSA patterns (Section 4.5.2, 4.5.3)?

5.1.1 Testbed. To evaluate our modeling framework, we constructed a consistent testbed combining county-level MRSA surveillance (2017–2024) with SDOH indicators and patient-day exposure data. All models were trained on 2017–2023 observations and tested on 2024, using MAE, RMSE, and R² as primary metrics. Categorical geographic identifiers were one-hot encoded, continuous variables were included directly, and lagged models removed first-year county entries lacking historical values. Each model was evaluated both for total county infections and for each hospital-type–specific subset. Rolling-origin cross-validation further assessed temporal robustness by training on expanding windows and forecasting the next year. This testbed provides a controlled and reproducible environment for answering our core scientific questions about MRSA risk across California.

5.2 Results, Observations, and Interpretation

5.2.1 Hospital Type Performance. Using the best-performing model (Random Forest with SDOH + Lag, tuned), we evaluated predictive accuracy across all nine hospital categories. Performance metrics for the 2024 test set are shown below, ordered by highest to lowest performance.

The model performs best in Major Teaching hospitals (R^2 = 0.93, MAE = 2.55) and Long-Term Acute Care facilities (R^2 = 0.90, MAE = 0.61). These settings treat large numbers of high-acuity patients and have long lengths of stay, which create steady, repeatable opportunities for MRSA transmission. Because the underlying epidemiological pressure is so consistent that the model can detect stable patterns year after year. Community Hospitals show moderate predictability. Community Large performs well (R^2 = 0.87, MAE = 0.85) and Community Small also performs strongly (R^2 = 0.82, MAE = 0.52), but Community Medium is more variable (R^2 = 0.51, MAE = 0.76). Medium facilities serve counties with fluctuating patient volumes and irregular introductions of MRSA, which makes the signal weaker.

Hospital Type	MAE	RMSE	R^2
Major Teaching	2.5473	4.5277	0.9303
Long-Term Acute Care	0.6065	1.7703	0.9029
Community Large	0.8549	2.1003	0.8666
Pediatric	0.1229	0.4164	0.8366
Community Small	0.5199	0.8727	0.8166
Rehabilitation Unit	0.0646	0.1640	0.6889
Community Medium	0.7624	2.0356	0.5057
Critical Access	0.0835	0.2586	0.0465
Free-Standing Rehabilitation Hospital	0.0026	0.0155	0.0000

Table 2: Performance of the RF + SDOH + Lag model across hospital types for the 2024 test set, sorted by R^2 from highest to lowest.

Lagged infection counts help stabilize these predictions by anchoring them to the previous year's burden. The lowest-performing groups — Pediatric (R^2 = 0.84, MAE = 0.12), Rehabilitation Units (R^2 = 0.69, MAE = 0.06), Critical Access Hospitals (R^2 = 0.05, MAE = 0.08), and Free-Standing Rehabilitation Hospitals ($R^2 \approx 0.00$, MAE = 0.0026) — all have very low case counts. When annual infections range from 0 to 1, the model can only capture so much: MRSA events occur as isolated, patient-specific incidents rather than as continuous transmission. In these situations, R^2 becomes unstable even when the absolute errors are tiny, because one unexpected case dramatically shifts the variance. Overall, these results show that predictability follows volume. High throughput facilities have enough consistent exposure to generate learnable patterns, while low-volume facilities experience MRSA as sporadic events, making them inherently unpredictable, no matter the model.

5.2.2 County Level MRSA patterns. As seen in Figure 1, MRSA infections are heavily concentrated in a small number of large, urban counties, including Los Angeles, San Diego, Sacramento, and Orange, which account for the majority of statewide burden from 2017 to 2024. These counties show strong temporal persistence: areas with high infection levels in one year almost always remain high the next, reflecting stable endemic transmission in densely populated regions with high healthcare utilization. In contrast, most rural counties report zero or near-zero infections. Their small populations produce sporadic, highly variable case counts, making MRSA patterns in these areas more stochastic than endemic. Overall, MRSA burden in California follows a clear urban–rural divide, driven by population scale, hospital exposure, and sustained transmission in large metropolitan counties.

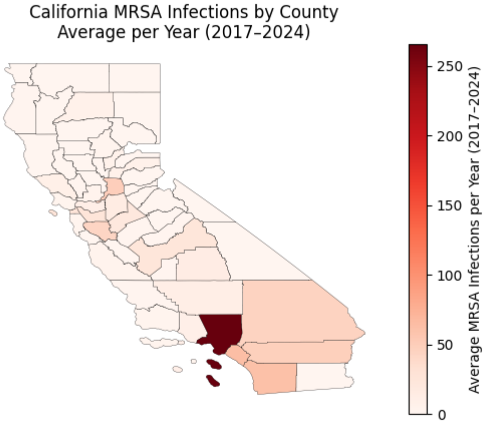
California MRSA Infections by County
Average per Year (2017–2024)
250
200
150
100
50
0
Average MRSA Infections per Year (2017–2024)

Figure 1: Average annual MRSA infections by county in California, 2017–2024

5.2.3 Feature Importance and Key Drivers of MRSA Risk. The tuned Random Forest model revealed a clear hierarchy of predictors that aligns closely with known drivers of MRSA transmission. As seen in table 3, Population-scale variables were the most influential: Labor Force (0.183) and Population / Population Rank (0.133 and 0.056) together accounted for nearly one-third of all model importance, reflecting the simple reality that larger, denser counties sustain more opportunities for MRSA spread. Measures of healthcare exposure were the next strongest group. Patient-days in Major Teaching hospitals (0.113), Community Medium hospitals (0.058), and Pediatric and Rehabilitation Unit settings (0.048 and 0.047) captured the intensity of inpatient contact networks where MRSA colonization typically occurs. Temporal predictors also ranked highly: lagged infections in Major Teaching (0.049) and Community Large (cl) hospitals (0.048) showed that previous-year MRSA burden is one of the strongest indicators of future burden, highlighting a persistent autoregressive structure. Finally, only a few SDOH variables contributed meaningfully. Most notably % Severe Housing Problems (0.034) and Income Ratio (0.003), suggesting that structural disadvantage plays a role but is secondary to population scale and healthcare exposure. Overall, the model indicates that MRSA risk in California counties is driven primarily by county size, intensity of inpatient care, and stable year-over-year transmission patterns.

5.2.4 Contribution of SDOH to Prediction Performance. The addition of Social Determinants of Health (SDOH) variables did not meaningfully improve MRSA prediction once population size and hospital utilization were included. As shown in Figure 2, models that incorporated SDOH alone (RF_with_SDOH) performed worse than the baseline Random Forest, both in MAE and R^2, indicating that SDOH did not add usable predictive signal on their own. Across all models, SDOH variables were highly collinear with one another and with population-scale measures such as county population, labor force, and patient-day exposure. Because these core features already capture the structural differences between counties, most SDOH indicators offered little new information. In linear regression, this multicollinearity caused model performance to collapse entirely, while in Random Forests the inclusion of SDOH generally reduced or failed to improve performance unless lagged infection counts were also present.

Rank	Feature	Importance	Variable Meaning
1	Labor Force	0.1830	Total county labor force (SDOH)
2	Population	0.1334	County population estimate
3	Patient-days: Major Teaching	0.1132	Bed-days in Major Teaching hospitals
4	Patient-days: Community Medium	0.0578	Bed-days in medium community hospitals
5	Population Rank	0.0561	County population percentile
6	Major Teaching infections (lag-1)	0.0487	Previous year's MRSA in Major Teaching hospitals
7	Community Large-Infections (lag-1)	0.0482	Previous year's MRSA in community large hospitals (cl)
8	Patient-days: Pediatric	0.0476	Bed-days in Pediatric hospitals
9	Patient-days: Rehabilitation Unit	0.0469	Bed-days in Rehabilitation Units
10	Patient-days: Community Large	0.0427	Bed-days in community large hospitals

Table 3: Top 10 Most Important Features in the Tuned Random Forest Model

Feature importance results confirmed that only a small subset of SDOH features (primarily % Severe Housing Problems and Income Ratio) contributed meaningful predictive value. These align with structural crowding and socioeconomic stressors, which are plausible mechanisms for elevated MRSA risk. Meanwhile, high-volatility variables such as PM2.5, % Smokers, and % Uninsured demonstrated near-zero importance, consistent with the noisy year-to-year fluctuations observed in the EDA. Overall, the patterns in Figure 2 make clear that SDOH provides limited incremental value: MRSA prediction is driven predominantly by population scale, healthcare utilization, and the persistence of prior-year infection levels, with only a few structural SDOH indicators capturing true epidemiologic risk.

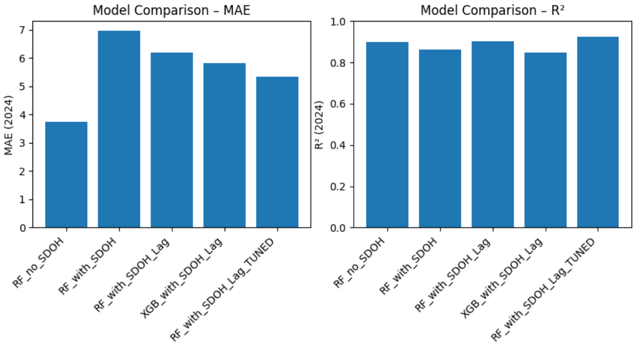

Figure 2: MAE and R^2 for all model configurations, comparing accuracy with and without SDOH and lag features

6 Conclusion and Future Work

We have demonstrated that county-level MRSA patterns in California are governed primarily by population scale, healthcare exposure, and temporal persistence. Across all models, the strongest predictor of future MRSA burden was the previous year's infections, confirming the highly autoregressive nature of MRSA transmission. The tuned Random Forest with SDOH + lag features achieved the best overall performance (RMSE = 9.02, R^2 = 0.924), outperforming linear, boosting, deep learning, and univariate time series approaches. Importantly, the hospital-type analysis revealed that model predictability directly tracks underlying epidemiology: large, high-acuity facilities such as Major Teaching and Long Term Acute Care hospitals (where MRSA transmission is consistent and sustained) were the easiest to model, whereas low-volume facilities exhibited inherently unstable patterns.

A second major conclusion is that Social Determinants of Health (SDOH) contributed limited incremental predictive power. Consistent with the EDA, only a few SDOH variables (e.g. % Severe Housing Problems, Income Ratio) meaningfully influenced risk, while most high-volatility indicators offered no consistent signal. This mirrors the structure revealed in feature importance: once population and patient-days are included, most SDOH variables become redundant or drowned out by noise. Rolling origin validation further confirmed that model stability improved across time, supporting the use of temporal features for MRSA forecasting.

Finally, although the predictive accuracy achieved is strong, the limited number of years (2017-2024) remains the largest barrier to model generalization. With only eight years of data, year to year fluctuations (especially those associated with COVID era utilization) can disproportionately influence learned patterns. Additional years would substantially strengthen both the temporal modeling and the ability to detect meaningful nonlinear SDOH effects.

6.1 Future Work

6.1.1 Forecasting. Future work for this model includes generating 2025 county-level MRSA forecasts, highlighting counties projected to have elevated infection burden. These forecasts would support targeted early surveillance, allowing hospitals in high-risk regions to prepare for increased MRSA pressure and potentially mitigate transmission through earlier intervention. Because the dataset spans only eight years, these predictions should be interpreted with caution. Limited temporal depth along with pandemic-era irregularities may constrain long-range forecasting accuracy. Additional years of data would substantially improve stability and reliability.

6.1.2 Epidemiological Applications. Future work could apply these findings to guide targeted MRSA surveillance, particularly focusing on the few SDOH indicators that consistently contributed meaningful signal, such as housing instability and income inequality. Incorporating these predictors into routine monitoring could support more informed policy decisions, hospital-level infection control planning, and prioritization of high-risk communities. The framework also opens the door for automated alert systems that use lag-adjusted trends to flag emerging hotspots or unexpected increases in MRSA burden.

6.1.3 Enhanced Temporal Modeling. Future work could incorporate multi year lag structures (e.g. 2-year and 3-year lags) to better capture long-term persistence in MRSA transmission patterns. This would allow the model to account for slower, structural trends that a single-year lag may miss. In addition, applying rolling averages or exponential smoothing to highly volatile SDOH variables could stabilize year-to-year noise and help reveal underlying socioeconomic signals that are currently masked by measurement variability.

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] Leo Breiman. 2001. Random Forests. 45, 1 (2001), 5–32. doi:10.1023/A:1010933404324

[4] 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.

[5] 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.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794. doi:10.1145/2939672.2939785

[7] François Chollet et al. 2015. Keras. https://keras.io.

[8] 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.

[9] 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.

[10] 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).

[11] 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.

[12] 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.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research , Volume 12, Pages 2825–2830* (2011).

[14] 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.

[15] Sima Siami-Namini and Akbar Siami Namin. 2018. Forecasting Economics and Financial Time Series: ARIMA vs. LSTM. arXiv:1803.06386 [cs.LG] https://arxiv.org/abs/1803.06386

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