SIR-SIS Fusion: MRSA and Influenza Co-infection Forecasting

Final Project Report: CSE-8803-EPI

Lauren Fogel
lfogel3@gatech.edu

Asma Khimani
akhimani30@gatech.edu

Harshini Vummadi
hvummadi3@gatech.edu

1 Introduction

In healthcare settings, preventing the spread of disease is a critical concern, especially in populations of highly vulnerable patients. Healthcare Associated Infections (HAIs) refers to infections that people can develop while receiving care for another health condition [U.S. Department of Health Human Services 2024]. In hospitals, 1 in 31 patients have at least one HAI at any given time [Centers for Disease Control and Prevention 2020]. Among HAIs, Methicillin-Resistant Staphylococcus aureus (MRSA) is a common infection that is associated with significant morbidity, mortality, length of hospital stay, and cost burden [Siddiqui and Koirala 2018], [Cosgrove et al. 2003], [Cosgrove et al. 2005]. Alongside this, seasonal outbreaks, such as influenza (flu), can further complicate infection control efforts, particularly when co-infection occurs between these viral and bacterial pathogens leading to worse patient outcomes [Liu et al. 2021]. However, hospitals do not always have effective ways to predict when and how these infections will spread concurrently. Therefore, it is important to understand the conditions that lead to co-infections, so hospitals can prevent these co-infection cases and improve patient outcomes.

1.1 Problem Statement

The problem this project attempts to solve is the lack of predictive models for understanding and assessing the risk of co-infections between MRSA and flu in healthcare settings. Co-infections have been shown to complicate patient conditions, with research indicating that 20.5-24.5% of patients with flu also experienced bacterial co-infections, particularly involving MRSA [Blyth et al. 2013]. These co-infections likely lead to more severe cases requiring intensive care, yet not much is known of the transmission dynamics between these viral and bacterial pathogens in a healthcare setting. Ultimately, this project addresses the need to better predict co-infection risks by developing a model to simulate the spread of both MRSA and flu within healthcare environments. By exploring how these infections interact, we aim to inform hospitals with insights they need to reduce the risk of co-infections.

2 Response to Milestone Comments

2.0.1 ***Dataset***. We originally received feedback about about how sufficient our 10 day real world dataset would be in our analysis. As a result, we decided to go with a combined approach where we used both a real world and synthetic dataset for our analysis. However, we discovered the real world dataset was insufficient at showing the full trends of the data and we felt it did not add anything meaningful to our analysis. Our solution was to adapt a single day from the real world dataset into a 30 day synthetic model that would allow us to rely on real world findings to expand the scope of our analysis and to capture a sufficient amount of trend data. Further details can be found in our data section.

2.0.2 ***Exploratory Data analysis***. We received feedback to conduct further exploratory analysis into our data so that we could understand it better and recognize patterns within the data. We generated summary statistics for the contact networks we created for MRSA and flu which we compared against each other. Additionally, we created plots to show the number of contacts per timestep, a useful metric that shows the patterns of movement in the data. Both are further discussed in the data section.

2.0.3 ***Calibrate the Model***. We received feedback about how we would decide our priors (*beta*, *gamma*, *alpha*, *delta*) in our SIR-SIS model. We decided the best approach to getting these values was to do a literature review and analyze the values being used in order to decide what we would make ours. Further details on the selection of priors can be found in the model and results sections.

3 Related Works

3.1 Infection Transmission in Hospitals

. To understand infection transmission in healthcare settings, we reviewed a paper that discusses the spread of MRSA in a hospital environment. The study focused on modeling the spread of MRSA in a dialysis unit at the University of Iowa Hospitals and Clinics. By using healthcare personnel movement data obtained from network sensors, the researchers built a contact network to conduct agent-based simulations to explore how MRSA transmission occurs within a unit. Specifically, they employed a SIS model to simulate the transmission of MRSA. This model is well suited to simulate MRSA infection dynamics because it shows that immunity is not conferred after infection. Furthermore, the researchers evaluated the architectural changes that could be made to the nurses' station to reduce to the spread of MRSA and found that increasing the surface area of the nurses' station lead to a reduction in MRSA infections ([Jang et al. 2019]. This shows how environmental factors and healthcare worker movement patterns can influence MRSA transmission. This study provides us with a methodological basis for modeling infection spread in healthcare settings using contact networks and a SIS model framework. By utilizing the researcher's data set, we aim to build a similar contact network that captures patient and healthcare worker interactions for simulations to model how co-infections between flu and MRSA may spread within a hospital setting. We plan to integrate a SIS model into our combined flu-MRSA model to allow for a realistic amount of MRSA reinfections to accurately simulate HAIs.

3.2 Co-infection Compartmental Models

One such paper discusses how two competing "memes" on two different social media platforms interact with each other, using differing metrics to simulate which "meme" will prevail [Wei et al. 2012]. In order to simulate the different social media platforms, they use a composite network where nodes are connected by two types of edges, one that connects the nodes within a platform and one that connects nodes between the two platforms. Within this composite network they simulate a SIS model with three compartments: susceptible population, population infected with meme 1, and population infected with meme 2. The paper uses a Non-Linear

Dynamic System (NLDS) to simulate how the infection spreads through each time step and calculates the eigenvalue which is used to determine which infection will prevail. The model is assessed on both real world and large scale simulation data. In smaller simulations we see that the meme with the larger starting eigenvalue always prevails and the other goes extinct, but in the larger scale data the meme with the smaller eigenvalue is still able to persist. They were also able to find if there was a small enough difference in the starting eigenvalues there would be no definite winner.

A strength of this paper is their usage of composite network, which is a novel method used to distinguish between populations and change how the infection spreads as a result. Given the rise in popularity of social media, the use of this approach gives a much more realistic picture of how information spreads through these systems. The use of eigenvalues also improves the paper as it condenses the topological information for better analysis of the dynamics between the two memes. The use of small and large scale data sets as well as real world data in their simulation also provides the results more robustness.

The paper could have benefited from further exploring the dynamics between the two infections. The infections are mutually exclusive which limits the scope of the predictions. A model where the viruses interact in more varied ways, such as having partial competition, co-infection, or contamination across layers, would provide addition metrics to enhance the data. Furthermore, the model does not account for infections on the same time step by assuming lower probability of infection and persistence or small time steps. As a result of this assumption, the interactivity of the two memes is low and would not be able to predict interactions between stronger viruses accurately.

. Another paper aims to improve the interactivity between concurrent contagions in a network [Stanoev et al. 2014]. The paper examines a discrete-time stochastic process for modeling the spread of multiple contagions in a network, represented by an undirected, connected graph with nodes and links. Each node can only be in one state at a time. Each node can change in one of two ways: a spontaneous transition to another state or a state change from contact. The spontaneous transition is similar to when a node recovers in an SIR model. They emphasize the possibility of multiple events occurring in the same time step and use approximations for the node probability to make the model numerically tractable. The model can be transformed into a deterministic system of difference and differential equations by assuming nodes' states are independent random variables. This leads to nonlinear dynamical systems that describe macroscopic spreading, using mean-field approximations. It explores a variety of different models, from the single contagion SIS to more complex multi-infection SIS models and rumor networks and extends them to work for multiple contagion networks.

The paper has many strengths and focuses on solving a key issue in other multiple contagion papers: handling multiple simultaneous infections within a single time step. Previous papers, avoid this problem by assuming either infinitesimally small time steps or independent infections. By using a discrete-time stochastic model, they are able to simplify the model and avoid this issue. Also, they utilize approximation methods which generalize the models well and showed high accuracy. To assess the accuracy of an approximation they compared the non-approximated and approximated versions of the deterministic counterpart. The predictions of fixed points are consistent between the approximated and non-approximated versions, even when the model gets more complex. This allows for further analysis of the parameter distribution through methods like Markov Chain Monte Carlo sampling for estimation. The probability mass vectors which are needed for the likelihood are efficiently computed using the approximation of the parameters. However, the model can not be generalized to many real world situations as many contagions do not fit into the models simplified schemes. Stochastic models are preferable for smaller populations and as a result, so this model may not work well for large data sets.

. Given our project focuses specifically on influenza and MRSA, we found it fitting to review a paper focused on the dynamics between two specific diseases. The following paper focuses on the effects of co infection between COVID-19 and kidney disease [Hye et al. 2024]. The goal of the paper was to understand the transmission dynamics between the two infections using a SIR model. They developed an SIR model with seven compartments for the various infection states that can occur. From the model simulations, they were able to identify equilibrium points for each disease separately and analyzed their stability based on their basic reproduction numbers which they calculated using a next-generation matrix. They also analyzed the infections together, where they found if the infection rate for either COVID-19 or kidney disease increases, the risk of people getting both diseases increases significantly.

The paper addresses a gap in research on how COVID-19 interacts with kidney disease, a relevant problem that has not been studied at the time. Comparing COVID-19 and kidney disease separately allows researchers to create a baseline of how the diseases behave separately. This design choice allowed them to better analyze the co-infection dynamics of COVID-19 and kidney infection. The paper also uses real-world data to validate their results.

The paper bases its parameters on real world data which cannot be done for every disease. It is optimized specifically to COVID-19 and kidney disease and therefore cannot be generalized like the previous models. While these weaknesses reduce the application potential of the paper, they do not affect the quality of the results.

3.3 Machine Learning Models

. In a 2024 study [Zhang et al. 2024], eight different machine learning models are compared, including Random Forest, Ada Boost, and Extreme Gradient Boosting (XGB), to predict co-infection in Lupus Nephritis (LN) cases. They showed how having LN could raise risk of co-infections due to suppressed immune system and decreased levels of white blood cells. Machine Learning (ML) models were used rather than mathematical models because ML models can better discern immune patterns and identify clusters to predict disease outcomes. Rather than looking at specific diseases for co-infection, the goal of the experiment was to predict the most likely infection locations. Respiratory infections were the most prevalent at 69.4%, and the models also predicted more bacterial co-infections than viral. This paper is unique because no other studies on LN co-infections have been done. They also keep things relatively simple by using easier implementable ML models found in packages like Scikit-learn [Pedregosa et al. 2011]. Using statistical analysis on

the outcomes of each model allowed them to determine the best performing model, which was XGB. However, the biggest weakness was the small sample size used. This limits the effectiveness of the model and could potentially lead to differing information and performance results for a larger sample.

. Another study conducted in 2024 compares the use of different statistical and ML models for predictive time series modeling of Tuberculosis (TB) and the Human Immunodeficiency Virus (HIV) co-infections [Abade et al. 2024]. There is a synergistic interaction between the two diseases, making viable treatments more complex and worsening clinical outcomes. While ML models were found to be the most accurate, the authors do concede that simpler models still provide useful guidance for managing TB/HIV co-infection.

This paper is very strong in many areas, including data processing, modeling, and evaluation metrics. There was extensive pre-processing done on the data, due to the size of the dataset and variability between reporting methods. Several modeling methods were used, with the best results coming from the custom Convolutional Neural Network and Long Short-Term Memory (CNN + LSTM) model and Bidirectional Long Short-Term Memory (LSTM) model. These ML models had the highest accuracy rates, and the best statistical model was Double Exponential Smoothing (DES) due to its efficiency. This study truly distinguishes itself due to the wide array of predictive models used, showcasing strength in utilizing cutting-edge data analysis techniques for application to intricate epidemiological problems.

Despite its many strengths, this study does show some weaknesses, mainly in the realm of timing issues and dataset constraints. While ML models are great for prediction, use of complex models like neural networks are time consuming and computationally expensive. The dataset used also may have some inaccuracies and missing data as mentioned earlier due to the nature of non-standard reporting practices across medical systems. The data spans 2013-2023, which means the data was affected by the COVID-19 pandemic. Late diagnoses and lesser treatments were common during this time, meaning the data was likely not accurate for this interval.

. These two papers both compare multiple ML models for the prediction of prevalence of co-infections for two different diseases. Interestingly, both tested Extreme Gradient Boosting. While Zhang et al. found this to be the best method for prediction, Abade et al. found it inferior to others because of sub-optimal configuration and potential overfitting. Each paper used different methods for analyzing their model output, which could also contribute to these differences, but this allows us to compare these methods to determine what will be best for our project.

4 Proposed Method

4.1 Intuition

Co-infection models are understudied. There are few examples of such models in literature, and none of these models focused on the spread of two infectious diseases at the same time, instead choosing to focus on one infectious disease and one non-infectious disease, such as an autoimmune disease. Most models of this type that do exist are also either mathematical models, or machine learning models. Our model is revolutionary not only due to its study of two infectious diseases, but also by incorporating both mathematical and machine learning models for a more insightful outcome.

4.2 Models and Techniques

4.2.1 Mathematical Model for Disease Spread. To simulate the co-infection dynamics of MRSA and flu, we have developed a combined SIR-SIS model. The SIR model will be used to represent the spread of flu because infected individuals can recover and gain a temporary immunity. The SIS model will be used to represent MRSA because individuals can become reinfected after recovery. Both of these models are integrated into a single combined model to allow for individuals to contract both infections simultaneously. This model allows us to study how the two infections interact with each other and how one infection may influence the transmission or severity of the other infection.

We decided on the parameters and compartments below to reflect each state of the model. Patients can be susceptible to both diseases, have one of or both diseases, or be recovered from flu, since it is modeled using the SIR model, but still infected or susceptible to MRSA. The parameters reflect each of these with transmission and recovery rates, plus an extra parameter to model how an individual who is currently infected with influenza is more susceptible to getting MRSA. We have only included this one parameter based on a study conducted by Keer Sun in 2014 [Sun and Metzger 2014], which showed that mice infected with flu had suppressed ability in macrophages and neutrophils to kill bacteria, making them more susceptible to bacterial infections. No studies were found which raised the likelihood of getting flu if one is infected with MRSA.

Parameters

- β_{flu}: Transmission rate of influenza
- β_{MRSA}: Transmission rate of MRSA
- γ_{flu}: Recovery rate of influenza
- γ_{MRSA}: Recovery rate of MRSA
- γ_{both}: Recovery rate from both diseases, taking in co-infection dynamics
- α: Co-infection penalty factor
- δ: Increased susceptibility to MRSA due to current infection with influenza

Model Compartments

- S: Susceptible to both influenza and MRSA
- I_{flu}: Infected with influenza only
- I_{MRSA}: Infected with MRSA only
- I_{both}: Co-infected with both influenza and MRSA
- $I_{\text{MRSA\_}R_{\text{flu}}}$: Recovered from influenza but infected with MRSA
- $S_{\text{MRSA\_}R_{\text{flu}}}$: Recovered from influenza but susceptible to MRSA

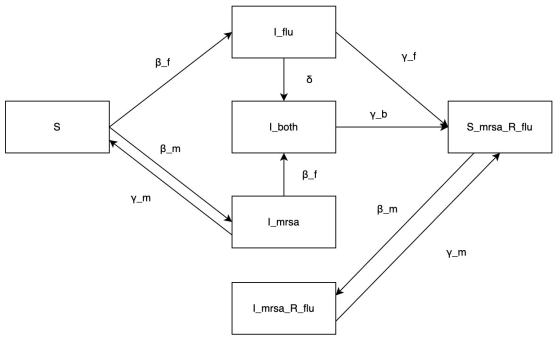

Figure 1: Flow Diagram of the Model

Differential Equations

Starting with an equation to represent the entire population:

$$N(t) = S(t)+I_{\text{flu}}(t)+I_{\text{MRSA}}(t)+I_{\text{both}}(t)+I_{\text{MRSA\_}R_{\text{flu}}}(t)+S_{\text{MRSA\_}R_{\text{flu}}}(t)$$

Then, calculate recovery rate of co-infected individuals:

$$\gamma_{\text{both}} = \frac{1}{\frac{1}{\gamma_{\text{flu}}} + \frac{1}{\gamma_{\text{MRSA}}} + \alpha}$$

The equations below are derived from the initial SIR/SIS model equations we discussed in class.

$$\frac{dS}{dt} = -\beta_{\text{flu}} S I_{\text{flu}} - \beta_{\text{MRSA}} S I_{\text{MRSA}}$$

$$\frac{dI_{\text{flu}}}{dt} = \beta_{\text{flu}} S I_{\text{flu}} - \gamma_{\text{flu}} I_{\text{flu}} - \delta\beta_{\text{MRSA}} I_{\text{flu}} I_{\text{MRSA}}$$

$$\frac{dI_{\text{MRSA}}}{dt} = \beta_{\text{MRSA}} S I_{\text{MRSA}} - \gamma_{\text{MRSA}} I_{\text{MRSA}} + \gamma_{\text{flu}} I_{\text{both}} - \delta\beta_{\text{flu}} I_{\text{MRSA}} I_{\text{flu}}$$

$$\frac{dI_{\text{both}}}{dt} = \delta\beta_{\text{MRSA}} I_{\text{flu}} I_{\text{MRSA}} + \delta\beta_{\text{flu}} I_{\text{MRSA}} I_{\text{flu}} - \gamma_{\text{both}} I_{\text{both}}$$

$$\frac{dI_{\text{MRSA\_}R_{\text{flu}}}}{dt} = \beta_{\text{MRSA}}(S_{\text{MRSA\_}R_{\text{flu}}})(I_{\text{MRSA}}) - \gamma_{\text{MRSA}}(I_{\text{MRSA\_}R_{\text{flu}}})$$

$$\frac{dS_{\text{MRSA\_}R_{\text{flu}}}}{dt} = -\beta_{\text{MRSA}}(S_{\text{MRSA\_}R_{\text{flu}}})(I_{\text{MRSA}}) + \gamma_{\text{flu}} I_{\text{flu}}$$

Assumptions Some assumptions were made for a reasonable experiment setup. No one can be infected with both MRSA and flu in the same time step. This is accomplished, following the competing meme paper [Wei et al. 2012], using small time steps so the likelihood of getting both diseases at once is negligible. Since we are concerned with modeling co-infection dynamics and not tracking an epidemic, we will not have a death state. We are also not setting a time limit on the recovered state for flu, since our data will not cover a long enough time period to have patients be re-susceptible. Contrary to this, a co-infected individual must recover from both diseases in the same time step; they cannot go back to only being infected with one disease, and must go into a recovered state. While there is no MRSA vaccine, flu vaccines are very common. Since our data tracks both dialysis patients and healthcare workers, we will assume that everyone is vaccinated for flu, since this is required for healthcare workers, and dialysis patients are heavily advised to get the vaccine due to being at higher risk for infection, according to the NIH [Sam et al. 2023]. This will manifest in a lower transmission rate of flu (β_{flu}). Finally, it is assumed if someone is infected with only one disease, they are susceptible to the other.

4.3 Dataset

4.3.1 ***Data Collection Process.*** The dataset we will use is Healthcare Personnel Movement Data [Jang et al. 2021], which is based on a network of sensors to track the movement of healthcare workers in a hospital setting, specifically a dialysis unit. The data set contains 10 days of the movements of healthcare workers and has fields such as badge ID, time, location of the nursing station, handwashing station and dialysis chairs, and time intervals of patients in and out of the dialysis chairs. The 4MB dataset will be downloaded from Kaggle and in the format of text and csv files.

Originally we were planning to use all 10 days of real world simulation to model our SIR-SIS model but quickly discovered that amount of days was insufficient for being able to capture the trends seen in the data. Therefore, we decided to use a synthetic network generated through the real world data so that our simulation had a foundation in real world data while having enough data points.

From the provided data, contact networks were generated for simulations by referring to a GitHub repository focused on COVID-19 transmission within a dialysis unit published the original authors of the dataset. We ran a script which was able to precompute contact networks based on proximity thresholds, spatial locations, and occupation of specific zones within the dialysis unit. The script was able to provide 30 days of simulated data for the most intensive day (Day 10) in the dialysis unit and includes two groups of patients on different schedules along with healthcare workers. The authors' original model for COVID-19 assumes close contact within a 6-foot distance for transmission, which follows closely with established flu transmission guidelines [U.S. Department of Veterans Affairs 2024], thus making it an ideal foundation. Further modifications were required to adapt the data for modeling MRSA transmission which has a contact distance of 1 foot and therefore has a much smaller contact network. We will generate contact networks between HCW's-patients and patient-patient based on the transmission requirements. Once the contact networks were computed we developed a script to read in the numpy input and convert it into a more readable csv format so that we could work with it easier. The contact csv includes information on day, timestep, and contacts that occurred during that timestep, one for flu and one for MRSA. We then created rows that showed the start and end of each contact so we could build our network graph edges.

Table 1: Summary Statistics

Metric	Mean (Flu)	Std Dev (Flu)	Max (Flu)	Mean (MRSA)	Std Dev (MRSA)	Max (MRSA)
Overall Degree	20.08	12.22	50.00	19.84	12.00	50.00
Overall Weighted Degree	44462.04	31111.35	130636.00	17916.55	5342.40	29362.00
HCW Degree	41.82	9.22	50.00	41.18	9.00	50.00
HCW Weighted Degree	95181.45	32023.85	130636.00	13541.45	4149.05	21675.00
Patient Degree	14.10	1.20	17.00	13.98	1.27	16.00
Patient Weighted Degree	30514.20	7078.44	49094.00	19119.70	4994.45	29362.00
HCW-Patient Edge Weight	1554.85	984.51	4415.00	288.95	178.48	867.00
Patient-Patient Edge Weight	3222.97	1915.22	9313.00	3217.42	1952.83	9452.00

4.3.2 ***Summary Statistics of Dataset.*** Table 1 shows the summary statistics of the dataset for flu transmission. Overall degree represents the number of unique connections an individual has in the network, while weighted degree accounts for the frequency of these interactions, where some individuals can have much more frequent or prolonged interactions. The contact type edge weights represents the strength of the interactions between healthcare workers and patients. These statistics show that patients have less contacts but longer durations of contact when compared to healthcare workers with patients, which provides insight on transmission potential.

Comparing the values between flu and MRSA we see that the unweighted degree values are all pretty similar across the board but the weighted degree values show there is a greater volume of flu contacts than there are for MRSA. This is an expected result as MRSA has a much smaller contact distance (1 foot) compared to the flu (6 feet

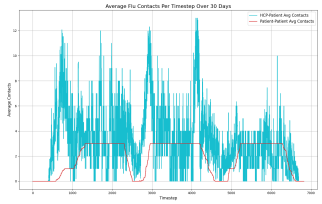

Figure 2: Average Flu Contacts Per Timestep Over 30 Days

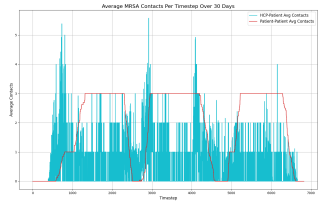

Figure 3: Average Flu Contacts Per Timestep Over 30 Days

4.3.3 ***Average Contacts Over Timestep Data.*** To better understand the movement trends in our dataset we decided to plot the average number of contacts individuals have for each timestep in our data. When looking at the plots we see results that strongly match what we expected to see given our knowledge of the dataset. Patients have prolonged contact with each other during dialysis sessions , which last a couple hours (Fig 2, 3). We see this represented by patient-patient contact which shows prolonged periods of contact three times a day. As for healthcare worker and patient contact, we see that it is greater than patient-patient contact which would make sense as we expect a greater frequency of interactions with healthcare workers and patients, even if these contacts are for smaller durations as we see with the fluctuations in the data. There also appear to be spikes in hcw-patient contacts at the beginning and end of what we believe to be the start and ends of dialysis sessions. When comparing flu contacts with MRSA contacts we see that patient-patient contact remains the same, suggesting the dialysis chairs are close enough to spread both infections (Fig 2, 3). The chairs may be a main source of infection spread due to the long duration of contact and proximity. Hcw-patient contacts are greatly reduced in MRSA but the trends seen in flu remain the same.

4.4 Model Implementation

We created a custom ODE model that models the dynamics of the co-infection between influenza and MRSA. Instead of using the equations above to calculate how nodes would move from state to state, adapted measures to represent more real-world data flow were implemented. For each possible state change, a new function was created. Infected neighbors were counted and a modified version of the equations above were used to calculate the probability of spread for each particular node. This value was run against a random number generator, and if it was less than the probability, then we could assume node infection. The new state of infection was returned if infected, or the original state was returned if no infection occurred. Recovery was more easily modeled, by checking a random number against each gamma value to see if a node recovered, and returning the proper recovery or re-susceptible state. To then call these functions, we have additional functions which call the appropriate scenario for the current state of each node, and then return all the updated states.

To actually model the disease spread, we build a graph with the people in our study represented as nodes. States are added for all nodes, and those initially infected are moved to infected states. The simulation is run day by day, first by iterating through each contact and adding the necessary edges for the daily graph, which are cleared and reset for the next day. Then, infection spread is run, states are updated, and finally, nodes are recovered. This output is then saved for the ML model, and graphed to visualize co-infection.

We also employ two helper functions for calculations. One calculates the recovery rate of co-infected individuals using the equation seen previously. The other was added to account for the way our model handles timestep data. Since each time step in our data represents 8 seconds, we had to find a way to account for this when applying our beta values. Gamma values for recovery are applied only once a day, as a person cannot recover multiple times a day. Beta values for transmission, however, can spread at any point of contact. So, we organized our data to get each new contact and organize them into rows, where each row is an occasion of contact. Based on the duration of this contact, a new beta value is calculated, which is then applied to the possibility of disease transmission.

$$\beta_{\text{new}} = \frac{\beta_{\text{initial}} * \text{duration}}{\text{number of time steps}}$$

Number of time steps is determined using a simple calculation of the number of time steps per minute, which at 8 seconds each is 7.5 time steps per minute, times the number of minutes considered for disease transmission. For example, if disease spreads after roughly 10 minutes, then $7.5 * 10 = 75$ time steps.

Multiple simulations are run for each set of conditions. Initial infected nodes are randomly chosen for each simulation to introduce randomness, and all random values will be redone as well. This creates different outcomes and disease spread for each model. These are all combined and averaged to get a smoother and more general output.

4.5 Machine Learning Models

In order to assess co-infection risk and identify under which conditions the risk is highest, we utilized ML algorithms, such as Logistic Regression (LR), Random Forest (RF), and XGB. This will be multiclass classification to predict the risk level of co-infection (high, medium, and low). We plan to use the simulation outputs from our custom SIR-SIS model as features in our ML models, such as prevalence of MRSA and flu, rate of change in MRSA and flu, and peak MRSA and flu prevalence.

4.5.1 Data Preprocessing. **Data Aggregation and Target Label Generation** To prepare the simulation data for model training and evaluation, we concatenated all simulations (550), where each row represented the state of the system at a specific day within a simulation, totaling 17,050 rows of data. The target variable, risk level,

was derived from the number of individuals co-infected with both flu and MRSA (I_both). Thresholds for categorizing risk levels were determined based on quantiles: Low Risk ($I_{\text{both}} \leq Q_{33}$), Medium Risk ($Q_{33} < I_{\text{both}} \leq Q_{66}$), High Risk ($I_{\text{both}} > Q_{66}$). **Feature Engineering** For the model to capture the dynamics of co-infection risk, the we created features derived from the simulations data. We engineered prevalence metrics, such as the proportion of individuals infected with MRSA (MRSA_prevalence) and the proportion of individuals infected with flu (flu_prevalence). These features provide insight into the overall spread of each infection at a given day. To capture temporal trends, we calculated the rate of change in prevalence for both infection, which is the day-to-day difference, where $\bar{X}$ is the prevalence metric:

$$\Delta X_t = X_t - X_{t-1}$$

Where X_t is the prevalence of the current day t, X_{t-1} is the prevalence at the previous day $t-1$, and ΔX_t is the change in prevalence between t and $t-1$.

To further capture temporal trends, binary indicators were created to flag peak prevalence events. The feature is_peak_flu identified whether flu prevalence reached its maximum value on a given day, while is_peak_MRSA indicated the same for MRSA prevalence. To account for early surges, we created is_first_peak_flu and is_first_peak_MRSA to flag the first occurrence of these peaks.

4.5.2 Model Development. **Formal Algorithm Descriptions** We evaluated three different supervised learning machine learning models, implementing each using Scikit-learn [Pedregosa et al. 2011]. **Logistic regression** is a statistical method that calculates the likelihood of an event occurring between two data factors. **Random Forest** is a machine learning algorithm that averages the outcome of multiple decision trees for classification and regression problems.**Gradient Boosting** builds upon each previous model to minimize overall prediction error. In each stage, regression trees are fit on the negative gradient of the loss function for optimization. **Scaling and Train-Test Split** The dataset was split into training (80%) and testing (20%) sets grouped by simulation. **Hyperparameter Optimization** To optimize the performance of the models, hyperparameter tuning was performed using a randomized search. The RF hyperparameters tuned included the number of estimators, maximum depth of the trees, minimum samples per split, and minimum samples per leaf. For LR, hyperparameters like regularization strength (C) and penalty type were tuned. XGBoost was tuned on parameters like the learning rate, maximum depth, and subsampling fraction.**Model Evaluation** Model performance was evaluated using metrics, such as prioritizing recall for the high-risk class and minimizing false negatives. When predicting co-infection in a healthcare setting, it is more harmful to misclassify high risk cases as medium or low risk when they are actually high risk (false negatives) than it is to misclassify low or medium risk cases as high risk (false positives). Although a false positive may cause unnecessary concern, a false negative, which misses a high risk co-infection scenario prediction, would be much more harmful because it can cause delayed intervention, which can lead to more co-infection spread and more harm.

5 Experiments and Results

5.1 Questions and Testbed

5.1.1 ***Testbed.*** All code was run using Microsoft VSCode and terminal on MacBooks with both Intel and M3 chips. Code was written in Python and Jupyter Notebooks, requiring various packages such as numpy, pandas, matplotlib, random, defaultdict, and networkx. ML models were implemented using Scikit-learn [Pedregosa et al. 2011]. Data was downloaded from https://www.kaggle.com/datasets/hankyujang/healthcare-personnel-movement-data, and then processed into csv files called by Python code. These are included in our code files, as they are relatively small.

5.1.2 ***Questions.*** The experiments in this project are designed to answer these questions:

- What factors contribute to the risk of co-infection of MRSA and flu?
- Can we accurately classify co-infection risk as low, medium, or high risk levels?
- Under which conditions is the risk of co-infection highest?
- What insights can we derive from the trained model and scenario based testing?

5.2 SIR-SIS Model

5.2.1 ***SIR and SIS Models.*** For this model, we first ran our simple SIR and SIS models on flu and MRSA respectively to get a baseline of how the infections spread without competition. We determined a transmission rate, or β value, of 0.3 for flu to use in all the models, as in this paper [Tan et al. 2013], they estimated a β of 0.411. We decided to use a lower value to model our assumption that all people the study are vaccinated, thus contributing less to disease spread. This same source chose a recovery rate for flu, or γ value, of 0.14. We felt this was too high, so we set our γ value to 0.1, since this was more reasonable and not recovering too many individuals too fast. For MRSA, we determined our β and γ values through a research paper [Beauparlant et al. 2016] which explored the metapopulation dynamics of MRSA spread through correctional facilities. They dynamically tested β and gave a range, 0.1 to 0.4, for the spread of MRSA. We used a beta value of 0.5 as MRSA transmission rates change throughout the year with rates peaking during the flu season so we adjusted our β accounting for that. We similarly adjusted their γ value from 0.3 to 0.4 to ensure we saw a reasonable amount of recovery for our smaller dataset.

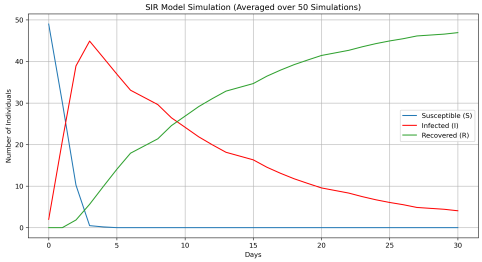

Figure 4: SIR Baseline Model

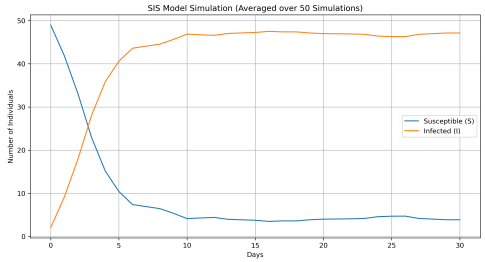

Figure 5: SIS Baseline Model

These models behave as expected for SIS and SIR models with the chosen parameters. While there are sharper peaks and divots than we might expect from a general SIR model, this is likely due to the nature of our data, which is quite different from what SIR models typically track. We can use these models later to compare infection rates without competition for both diseases.

5.2.2 ***Co-Infection Model.*** We ran the full co-infection model with our optimized parameters from the SIS and SIR models. Additionally, we chose a δ value of 0.6 based on this paper [Sun and Metzger 2014] which found that having influenza increased MRSA susceptibility by $\frac{3}{5}$. Our co-infection penalty factor, α, was chosen by testing several values before deciding that it should take an extra 10% of total infected time to recover from two diseases at once. The full model was run for 500 simulations to get the best idea of how the model performs, as well as common co-infection dynamics.

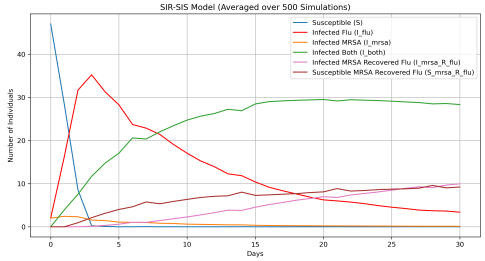

Figure 6: SIR-SIS Model, Initial 500 Simulation Run

Comparing to the initial SIS and SIR models, it is clear that competition between two infectious diseases significantly decreases their initial spread. MRSA only infection values stay very small for the duration, while flu only infection peaks early and then decreases as co-infection rises. Significant co-infection levels in this model shows the threat of opportunistic pathogens such as MRSA in a hospital setting when combined with another infectious disease like flu. While in the SIR model the flu recovered population is almost everyone by the end of the 30 days, in the co-infection model the I_both population continues to keep people infected with the flu. The decreased recovery rate has devastating impacts on the dialysis unit and shows the importance of monitoring co-infection rates in such a setting.

5.2.3 ***Evaluate Conditions.*** We then decided to test our model and co-infection dynamics by changing certain aspects of the model. We decided to focus on 4 aspects of our model: number of initial nodes, type of node, different contact times and different deltas.

We started by changing the number of initially infected nodes from two of each disease, one HCW and one patient each, to 6 flu and 2 MRSA, and then 6 MRSA and 2 flu. These results were not much different than our initial results, likely due to our randomization of initial infected nodes, as well as co-infection dynamics and contact relationships between nodes.

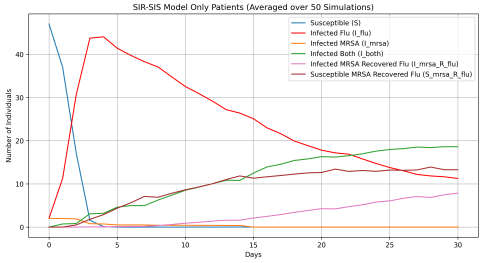

Figure 7: SIR-SIS Model, Only Patients

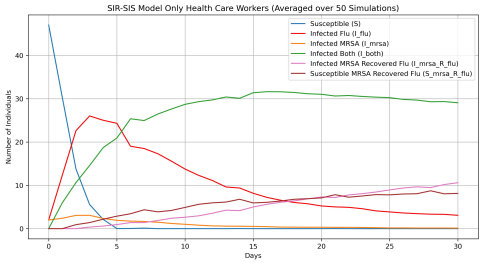

Figure 8: SIR-SIS Model, Only Healthcare Workers

Given that number of nodes seemed to make little difference on our model, we decided to focus on the type of node instead. We created two conditions: only patients and only healthcare workers to see if the type of node makes an impact on spread. We theorized that two types of nodes would show different spread since they displayed different patterns in the preliminary analysis of the data we did. Looking at the results (Fig 7, 8) we see a pretty drastic difference the two. For the patient only plot, spread of flu is much higher than in our base plot, MRSA struggles to spread on it's own, and co-infection barely spreads. The results support the notion that patients have limited but prolonged contact with one another, which we saw in Fig 2. Infected flu spreads faster since the contacts between patients are much stronger and MRSA struggles to co-infect with the limited number of contacts. On the other hand, we see in the healthcare worker only plot (Fig 8) that flu has a more difficult time spreading independently while co-infection remains relatively unaffected from base conditions. Given that healthcare workers have more contacts, even if said contacts are shorter, we can see how MRSA would continue to spread with the wider access to people that it has. It could be due to the increased presence of MRSA that flu struggles to compete and displays a much lower independent rate of infection, or the lack of longer patient contacts.

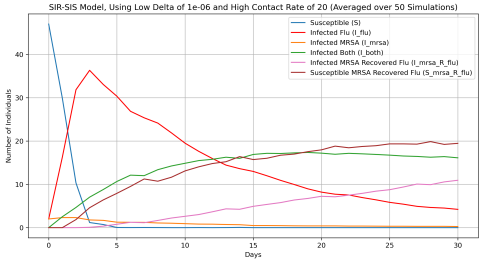

Figure 9: SIR-SIS Model, Low Co-Infection Spread

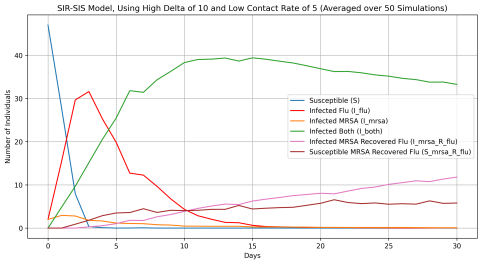

Figure 10: SIR-SIS Model, High Co-Infection Spread

For the last two conditions we decided to combine our variables after running them separately to create conditions for the optimal low and high co-infection states. For the time to spread infection our default value was 10 minutes, which is the time contact to have the full beta value chance of catching or spreading the disease. For this variable, we changed contact time to 5 and 20 minutes, to see how different levels of infection spread time effected the model. Additionally, we used *delta* as a metric to see how the intensity of co-infection effects the results. We heavily modified *delta* in our conditions (1e-6 vs 10) to see the full impact of the variable. After plotting these conditions together, we selected the plots with the highest and lowest co infection spread to evaluate (Fig 9, 10).

Looking at our low co-infection plot (Fig 9), we see that co-infection stays relatively low compared to our base model. With a reduced chance to spread (20 mins) and an almost non-existent *delta* (1e-6), the diseases have a difficult time spreading due to overall lower beta values. Flu infection values remain high at the beginning of the plot, suggesting that lower beta values for MRSA might have helped it compete better or that it spreads better through long term contact. Overall, our low co-infection graph shows the importance of reduced contact risk through social distancing measures and the impact MRSA has on flu spread.

The high co-infection plot (Fig 10), shows co-infection at a higher rate than our base graph, where we see almost complete domination of I_both. Given an overall easier chance to spread (5 mins) and a boost to *beta*_flu through an increased *delta* (10) we understand why we see an increased presence of co-infected nodes. The high co-infection graph shows the risks of high contact and the susceptibility of infections to each other.

5.3 Machine Learning

Model Performance The RF model achieved the highest AUC of 0.99 and a strong recall for the high risk class of 0.99, with 8 false negatives in this class. LR achieved lower performance an AUC of 0.87 and 0.18 recall for the high risk class with 806 false negatives. XGB performed well with an AUC of 0.99 and recall of 0.98 for the high risk class, with 22 false negatives. The RF model was selected for its superior recall metrics and minimal false negatives. **Feature Importance** The most important features contributing to the model's predictions were prevalence of MRSA with an importance score of 0.72 and prevalence of flu with an importance score of 0.1612, followed by the rate of change metrics. Peak related features, were not influential but they still provide additional context of the dynamics of infection spread. This highlights that prevalence and trends in infection growth are key factors of co-infection risk.

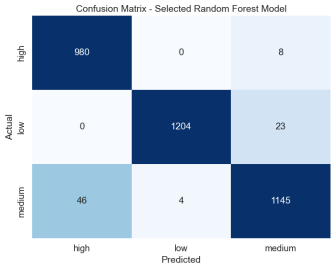

Figure 11: Confusion Matrix

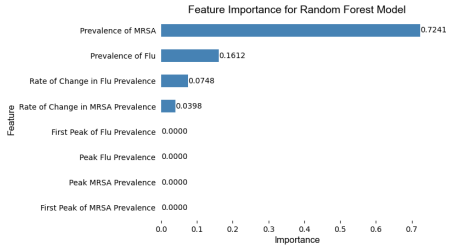

Figure 12: Feature Importance Plot

Scenario-Based Risk Evaluation To validate the model's predictions and provide actionable insights, we generated scenarios by varying important features, prevalence of MRSA and prevalence of flu across realistic ranges observed in the simulation data. The RF model identified 16 high-risk scenarios, primarily associated with high flu prevalence combined with moderate MRSA prevalence. These scenarios consistently showed high-risk probabilities exceeding 72%, which suggests that simultaneous surges in flu and MRSA prevalence significantly increase the likelihood of co-infection.

6 Conclusion and Future Directions

Our model shows the benefits and drawbacks of using both ODE co-infection model and machine learning to model and evaluate co-infection dynamics of flu and MRSA. Many factors and parameters affect co-infection rates in the SIR-SIS model, as is evidenced by the outcome above. In future explorations, we may want to add an exposed state to the ODE model and explore other methods such as an agent based model. Modifying certain parameters leads to findings that show the significant variance in co-infection rates. Our findings highlight the importance of monitoring increases in both flu and MRSA prevalence to identify high-risk scenarios of co-infection. The model's most important features, prevalence of MRSA and flu, suggests that real-time infection monitoring could help mitigate co-infection risk in healthcare settings. Also, the scenario-based risk analysis provides valuable insights and a potential resource for hospitals to evaluate conditions under which co-infection risk is highest allowing for timely intervention. **Contributions** All group members contributed equally to the project. Harshini did primary data processing and some ODE model implementation, Lauren did some data processing and primary ODE model implementation, and Asma handled the ML model.

References

André Abade, Lucas Faria Porto, Alessandro Rolim Scholze, Daniely Kuntath, Nathan da Silva Barros, Thaís Zamboni Berra, Antonio Carlos Vieira Ramos, Ricardo Alexandre Arcêncio, and Josilene Dália Alves. 2024. A comparative analysis of classical and machine learning methods for forecasting TB/HIV co-infection. *Scientific Reports* 14, 1 (2024), 18991.

Marc Beauparlant et al. 2016. A metapopulation model for the spread of MRSA in correctional facilities. *Infectious Disease Modelling* 1, 1 (2016), 11–27.

Christopher C Blyth, Steve AR Webb, Jen Kok, Dominic E Dwyer, Sebastiaan J van Hal, Hong Foo, Andrew N Ginn, Alison M Kesson, Ian Seppelt, Jonathan R Iredell, et al. 2013. The impact of bacterial and viral co-infection in severe influenza. *Influenza and other respiratory viruses* 7, 2 (2013), 168–176.

Centers for Disease Control and Prevention. 2020. Healthcare-Associated Infections Data Portal. https://www.cdc.gov/hai/data/portal/index.html. Retrieved from external site on health.gov.

Sara E Cosgrove, Youlin Qi, Keith S Kaye, Stephan Harbarth, Adolf W Karchmer, and Yehuda Carmeli. 2005. The impact of methicillin resistance in Staphylococcus aureus bacteremia on patient outcomes: mortality, length of stay, and hospital charges. *Infection Control & Hospital Epidemiology* 26, 2 (2005), 166–174.

Sara E Cosgrove, George Sakoulas, Eli N Perencevich, Mitchell J Schwaber, Adolf W Karchmer, and Yehuda Carmeli. 2003. Comparison of mortality associated with methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteremia: a meta-analysis. *Clinical infectious diseases* 36, 1 (2003), 53–59.

Md Abdul Hye, Md Haider Ali Biswas, Mohammed Forhad Uddin, and Md M Rahman. 2024. A mathematical model for the transmission of co-infection with COVID-19 and kidney disease. *Scientific Reports* 14, 1 (2024), 5680.

Hankyu Jang, Samuel Justice, Philip M Polgreen, Alberto M Segre, Daniel K Sewell, and Sriram V Pemmaraju. 2019. Evaluating architectural changes to alter pathogen dynamics in a dialysis unit: for the CDC MInD-healthcare group. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 961–968.

Hankyu Jang, Philip M. Polgreen, Alberto M. Segre, and Sriram V. Pemmaraju. 2021. COVID-19 modeling and non-pharmaceutical interventions in an outpatient dialysis unit. *PLoS Computational Biology* (2021).

Yingzhi Liu, Lowell Ling, Sunny H Wong, Maggie HT Wang, J Ross Fitzgerald, Xuan Zou, Shisong Fang, Xiaodong Liu, Xiansong Wang, Wei Hu, et al. 2021. Outcomes of respiratory viral-bacterial co-infection in adult hospitalized patients. *EClinicalMedicine* 37 (2021).

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* 12 (2011), 2825–2830.

Ramin Sam, Laura Rankin, Ifeoma Ulasi, Luc Frantzen, Dorothea Nitsch, David Henner, Donald Molony, John Wagner, Jing Chen, Sanjay Kumar Agarwal, et al. 2023. Vaccination for Patients Receiving Dialysis. *Kidney Medicine* (2023), 100775.

Abdul H Siddiqui and Janak Koirala. 2018. Methicillin resistant Staphylococcus aureus. (2018).

Angel Stanoev, Daniel Trpevski, and Ljupco Kocarev. 2014. Modeling the spread of multiple concurrent contagions on networks. *PloS One* 9, 6 (2014), e95669.

Keer Sun and Dennis W Metzger. 2014. Influenza infection suppresses NADPH oxidase–dependent phagocytic bacterial clearance and enhances susceptibility to secondary methicillin-resistant Staphylococcus aureus infection. *The Journal of Immunology* 192, 7 (2014), 3301–3307.

Xuhui Tan, Lingling Yuan, Jingjing Zhou, Yinan Zheng, and Fen Yang. 2013. Modeling the initial transmission dynamics of influenza A H1N1 in Guangdong Province, China. *International Journal of Infectious Diseases* 17, 7 (2013), e479–e484.

U.S. Department of Health Human Services. 2024. Health Care-Associated Infections. https://www.hhs.gov/oidp/topics/health-care-associated-infections/index.html. Accessed: 2024-10-08.

U.S. Department of Veterans Affairs. 2024. About Flu - National Center for Health Promotion and Disease Prevention. https://www.prevention.va.gov/flu/about.asp Accessed: 2024-11-05.

Xuetao Wei, Nicholas Valler, B Aditya Prakash, Iulian Neamtiu, Michalis Faloutsos, and Christos Faloutsos. 2012. Competing memes propagation on networks: a case study of composite networks. *ACM SIGCOMM Computer Communication Review* 42, 5 (2012), 5–12.

Jiaqian Zhang, Bo Chen, Jiu Liu, Pengfei Chai, Hongjiang Liu, Yuehong Chen, Huan Liu, Geng Yin, Shengxiao Zhang, Caihong Wang, et al. 2024. Predictive modeling of co-infection in lupus nephritis using multiple machine learning algorithms. *Scientific Reports* 14, 1 (2024), 9242.