Predicting the Impact and Trends of SARS-CoV-2 on the Respiratory Viral Season in Pittsburgh Using Interpretable Machine Learning Forecast Models: A Quality Improvement (QI) Retrospective Study

by

Shikha Puri

Bachelor of Science in Biological Science, University of Pittsburgh, 2020

Submitted to the Graduate Faculty of the

Department of Infectious Diseases and Microbiology

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Public Health

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This essay is submitted

by

Shikha Puri

on

April 25, 2024

and approved by

Essay Advisor: Jeremy Martinson, DPhil, Assistant Professor and Program Director, Infectious Diseases and Microbiology and Human Genetics, School of Public Health, University of Pittsburgh

Essay Reader: William Pasculle, ScD, Associate Professor Emeritus, University of Pittsburgh Medical Center

Predicting the Impact and Trends of SARS-CoV-2 on the Respiratory Viral Season in Pittsburgh Using Interpretable Machine Learning Forecast Models: A Quality Improvement (QI) Retrospective Study

Shikha Puri, MPH

University of Pittsburgh, 2024

Abstract

This Quality Improvement (QI) project utilizes predictive modeling to understand the dynamics of the COVID-19 pandemic, particularly examining the interaction with the respiratory virus season (RVS) encompassing Respiratory Syncytial Virus (RSV), Influenza, and SARS-CoV-2. This project seeks to determine whether COVID-19 will remain an additional burden on laboratories or diminish, making it another respiratory virus in the RVS. The analysis is from October 2015 to December 2023, examining incidence and ICD-10 cases from UPMC Shadyside and Presbyterian hospitals in Pittsburgh. This analysis compared pre- and post-COVID-19 periods, revealing evolving burdens on laboratories and hospitals. Our exploratory data analysis (EDA) visualizes the seasonal trends of the respiratory viruses, highlighting a shift in typical RVS patterns coinciding with the onset of SARS-CoV-2. Simple and Seasonal Naïve forecasting models provide baseline insights, while ARIMA and SARIMA models offer more advanced prediction techniques, acknowledging data complexities post-COVID-19. Despite SARIMA's superior performance, challenges arise due to limited post-pandemic data, emphasizing the need for continued data collection. The public health implications for our research are for proactive healthcare planning and understanding COVID-19's trajectory as a potentially endemic virus. Future endeavors will focus on continued data collection to refine the predictive models, create effective resource allocation strategies, and relieve the healthcare burden for potential future pandemics.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

The COVID-19 pandemic in 2020 has had a profound impact on global health, economies, research, and public health. Since its emergence in late 2019, COVID-19 spread rapidly, leading to more than 600 million people infected and over 6 million deaths reported deaths worldwide including 1.2 million deaths in the United States (JHU CSSE COVID-19 data up to March 10, 2023) [1]. The resulting impact of the pandemic strained the healthcare systems in many countries and led to a shortage of resources, disrupted economies, and altered daily life for people everywhere in the world [2].

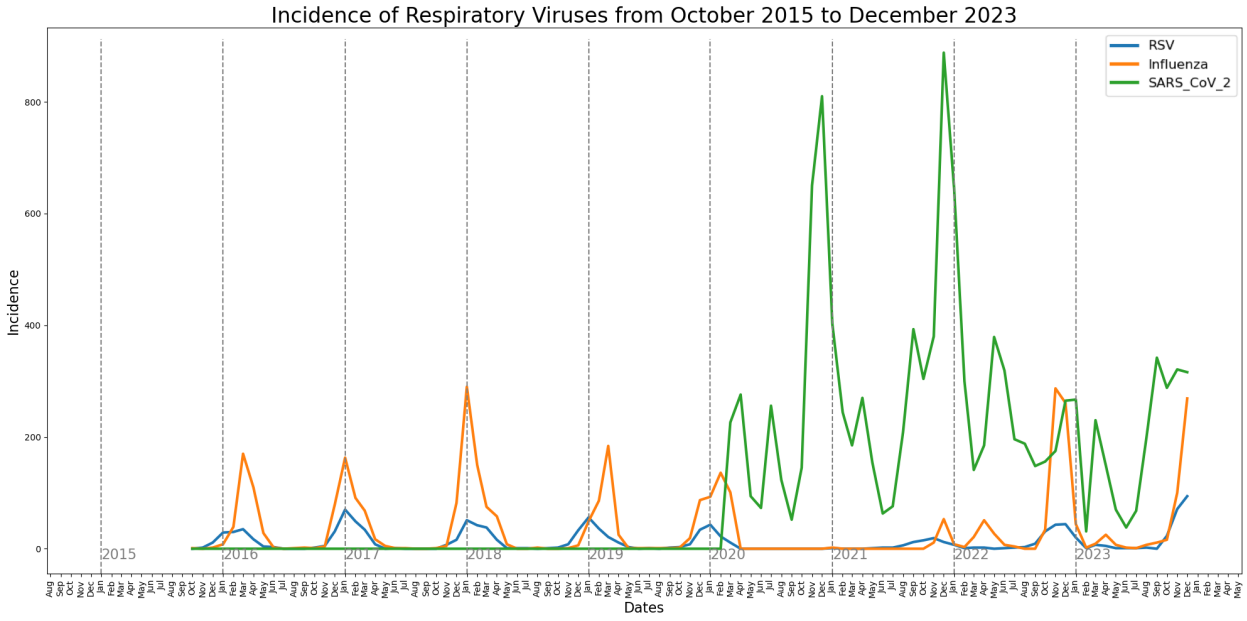


Figure 1. Incidence of Respiratory Viruses from October 2015 to December 2023.

Monthly incidence of respiratory virus infections: RSV (blue), Influenza (orange), and SARS-CoV-2 (green) from October 2015 to December 2023 using laboratory respiratory viral results from UPMC.

One of the unknown challenges posed by COVID-19 is its effect and impact on the incidence of respiratory viral infections, the so-called respiratory virus season (RVS). The RVS typically occurs from November to April, and it is characterized by a peak in respiratory virus infections, including Influenza, Respiratory Syncytial Virus (RSV), and Rhinovirus. Initially during the start of the pandemic, the concern was that SARS-CoV-2 would co-circulate with other major respiratory viruses like Influenza and overwhelm the healthcare systems. Figure 1 illustrates the incidence of the three respiratory viruses from 2015 to 2023. The sudden absence in cases of Influenza and RSV following the initial onset of COVID-19 highlights the potential shift in typical seasonal patterns and posing new challenges in public health.

The pandemic has led to an increased demand for healthcare services and resources, including ICU beds, ventilators, medical personnel, and testing. Both hospitals and laboratories faced challenges in managing the influx of COVID-19 patients, while continuing to provide care for other major illnesses. However, hospitals were forced to divide resources and limited staff between hospitalized COVID-19 patients and non-COVID patients. Along with the staff and supply shortages, the hospitals dealt with financial struggles caused by increased expenses and a reduction of non-COVID procedures [3]. In the initial months of the pandemic, between March 1, 2020 and June 30, 2020, US hospitals reported a total loss of over $200 billion because of a 45% decrease in usual treatment and operating revenue; this is also including the relief funds of $175 billion provided by the CARES Act and PPP Provider Relief Fund [4]. This financial strain was calculated by comparing the year before the pandemic (2019) with the first two years of the pandemic [5]. The strain on healthcare systems has highlighted the need for accurate prediction models to anticipate the future impact of COVID-19 on the RVS and future pandemics to effectively allocate resources.

The 1918 influenza pandemic infected one-third of the global population and caused an estimated 50 million deaths; however, it eventually became a seasonal flu virus that continues to circulate today [6]. To understand the trajectory of COVID-19 and its potential to become part of the seasonal respiratory viruses, it is important to consider the lessons from past pandemics. The 1918 influenza pandemic, caused by the H1N1 influenza virus, provides insight into the dynamics of pandemics, evolution of viruses in a population, and how it becomes part of the regular RVS [6].

By comparing the epidemiological and viral patterns of past pandemics with the recent burdens of COVID-19, we can gain insights into the future course of COVID-19. Understanding whether COVID-19 will remain an additional burden on healthcare systems or diminish its severity and frequency to integrate into the RVS is crucial for future planning and resource allocation.

Predicting the future trends of COVID-19 and the RVS requires a comprehensive analysis of epidemiological data, including the incidence of respiratory viruses from laboratory viral tests and hospital ICD 10 data. Machine learning models, such as the time series forecasting models used in this study, offer a powerful tool to analyze complex data and predict future trends. In the context of the RVS, time series models are effective in utilizing temporal and seasonal patterns to make predictions. The models used in the study are Simple Naïve Forecasting, Seasonal Naïve Forecasting, Autoregressive Integrated Moving Average (ARIMA), and Seasonal Autoregressive Integrated Moving-Average (SARIMA).

Each model has its own advantage and unique perspective on forecasting predictions. Naïve Forecasting (NF) models are the simplest models in time series analysis, and it assumes that the future value of a variable will be equal to the last observed value. ARIMA models uses current observations along with lagged observations to capture the autocorrelation structure of the full

time series data and captures non-stationary data trends, but mainly for linear-time series analysis [7]. Using three different tuning parameters, ARIMA can capture both the autocorrelation and trend in the time series data [7]. SARIMA incorporates seasonality into the ARIMA models as an additional parameter, allowing it to capture the cyclic fluctuations at fixed intervals [8]. Each model has its own strengths and can provide a comprehensive analysis of the progression and implications of COVID-19 on the RVS, to develop strategies for healthcare systems to be better prepared for future pandemics.

2.0 Methods

2.1 Study Design

This study uses a retrospective cohort design to analyze the impact of COVID-19 on the respiratory virus season (RVS) by looking at the healthcare and laboratory burdens. We collected and analyzed two datasets, the laboratory respiratory viral tests (Incidence) and hospital cases (ICD-10 codes), from October 2015 to December 2023.

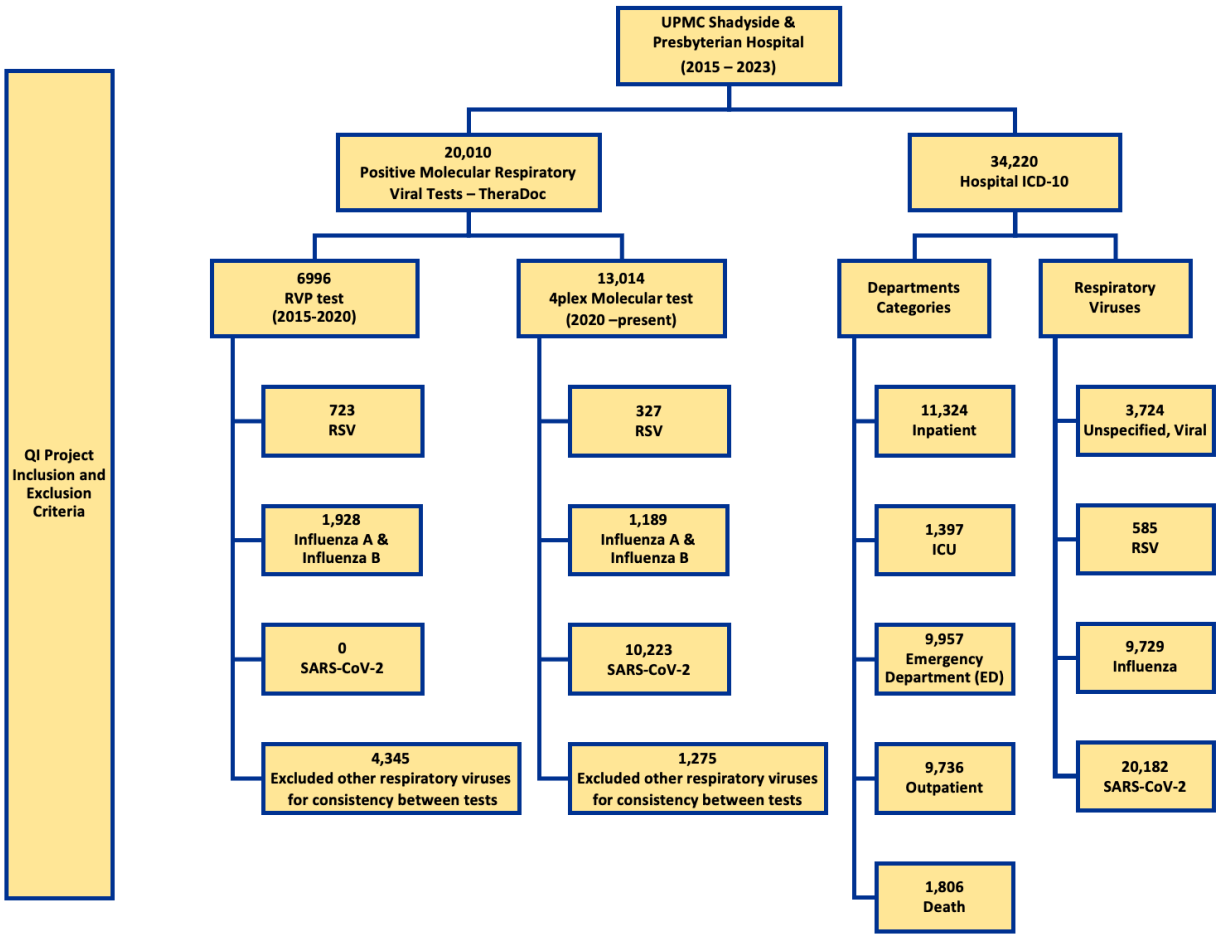


Figure 2. Inclusion and Exclusion Criteria Flow Diagram.

2.1.1 Laboratory Respiratory Viral Tests

Twenty thousand laboratory respiratory viral test results were collected from TheraDoc (Premier, Salt Lake City, UT), an electronic decision support application. Tests from 2015 to 2019 used the Respiratory Viral Panel (RVP) (ePlex© Respiratory Pathogen Panel, Roche diagnostics, Indianapolis, IN), which detects over 20 respiratory viruses and their subtypes in a single test [9]. Tests from 2020-2023 used the Gene-ExpertXpert© Xpress CoV-2/Flu/RSV plus Molecular test (Cepheid, Sunnyvale, CA) (4-plex), which only looks at the primary four viruses: RSV, Influenza

A and B, and SARS-CoV-2. Only the laboratory results for Respiratory Syncytial Virus (RSV), Influenza (A and B combined), and SARS-CoV-2 were included in this study to keep the viruses consistent through the years.

2.1.2 ICD-10 Codes

The second dataset is the viral respiratory infections ICD-10 codes obtained from UPMC Presbyterian and Shadyside hospitals from October 2015 to December 2023. The ICD-10 codes or cases are categorized into five categories: Inpatient, ICU, Emergency Department (ED), Outpatient, and Death. Comparing the number of cases in each category throughout the years will quantify the hospital burden and illustrate the impact of COVID-19 on the RVS. Because multiple ICD-10 codes denote each virus, we combined each into a separate single category for the analysis. In the study, ICD-10 codes are categorized into Influenza, RSV, SARS-CoV-2, and Unspecified Viral (which includes secondary respiratory illnesses like pneumonia or bronchitis that can arise as complications from these three respiratory viruses).

2.1.3 Data Analysis Methods

All statistical analyses and calculations were performed using Excel, R software and python. The two datasets— respiratory viral incidence and ICD-10 codes — were analyzed using classical statistical methods and machine learning predictive modeling techniques. Before creating the predictive models, we looked at the original data using Exploratory Data Analysis (EDA) to visualize the time series data in chronological sequence and observe any obvious trends. To prepare the data for model fitting, the two datasets were preprocessed to enhance data reliability

and interpretability. This includes the removal of missing data, ensuring consistency across all variables, and normalizing or standardizing the data to facilitate accurate modeling. The datasets were analyzed using four time series forecasting models —Simple and Seasonal NF, ARIMA, and SARIMA models. These models were used because of their effectiveness in capturing temporal and seasonal patterns and making predictions, in the context of the RVS [7]. To assess the performance of each model, various metrics, such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) were compared. These metrics provided the accuracy and the best performing models in forecasting trends in RVS incidence and ICD-10 codes [7]. After the best model selection, both datasets were used to generate forecasts for the future trends in RVS. These steps are used to create and evaluate the predictive trends for incidence and ICD-10 codes to assess the future of COVID-19 and the RVS in hospitals and laboratories.

2.1.4 Ethical Considerations

This study was approved by the Quality Improvement (QI) Committee of UPMC. This study adheres to the ethical guidelines for research involving healthcare and patient information. Data were anonymized and stored securely to protect patient confidentiality.

3.0 Results

3.1 Exploratory Data Analysis (EDA)

This study first conducted an EDA of respiratory virus incidence and ICD-10 cases from October 2015 to December 2023. The EDA includes various graphical visualizations to understand the characteristics of a time series dataset before applying any modeling techniques. The data is visualized in its original form before normalization to map out the chronological sequence of data and to analyze any obvious trends and seasonal variations.

3.1.1 Incidence EDA

The analysis of respiratory virus incidence from October 2015 to December 2023 (Fig. 3a) visualizes the annual trends for each virus. Pre-2020, RSV and influenza exhibited periodic peaks, aligning with typical respiratory virus seasons, which were particularly noticeable from November to March. SARS-CoV-2 suddenly appeared in 2020, indicating the onset of the COVID-19 pandemic. The association of SARS-CoV-2 on the RVS dynamic was evident by the decline in the incidence of the other respiratory viruses. Figure 3b narrows the focus to the period between 2020 and 2023 to display the emergence of COVID-19 and the RVS. The notable absence of incidence for RSV and Influenza from March 2020 to August 2021, showcase the drastic shift in normal RSV patterns from 2020 onwards.

Comparing the annual trends pre- and post- SARS-CoV-2 from 2015-2019 (Fig. 4a) versus 2020-2023 (Fig. 4b) revealed the change in RVS dynamics. The annual trend of incidence for

respiratory viruses uses a different perspective of time series to offer insight into monthly trends from 2015 to 2023. By eliminating seasonal variations, the plots assess each virus's incidence across a specific month over the years. For example, these plots show if the incidence of Influenza in January increased, decreased, or remained constant from 2015 to 2023. Pre-COVID-19, SARS-CoV-2 was negligible, contrasting with the post-2020 with increasing incidence trends from 2020 to 2021. After 2022, the SARS-CoV-2 incidence trends are decreasing. Influenza, compared to 2021 to 2022, there was a sizeable increasing trend for incidence in Influenza, which is from the absence of Influenza and RSV incidence from March 2020 to August 2021. Influenza and RSV show varying patterns, with noticeable changes in incidence during the pandemic, indicating a change in the normal dynamics of respiratory viruses. These incidence EDAs provide insight into understanding the epidemiological characteristics of RVS patterns and the association of the COVID-19 pandemic.

3.1.2 ICD 10 EDA

The exploratory data analysis of ICD-10 cases was conducted on respiratory virus hospital cases from October 2015 to December 2023. The analysis focused on two main subsets of data: ICD-10 cases of respiratory virus cases and ICD-10 cases across hospital departments.

The breakdown of ICD-10 respiratory virus cases by virus type (Fig. 5a) shows that Influenza and RSV dominated the ICD-10 cases until 2020. In early 2020, there was a significant spike in Unspecified Viral cases during the onset of COVID-19, but by 2021, these unspecified viral cases disappeared. The spikes in SARS-CoV-2 ICD-10 cases closely mirror the incidence plot of respiratory viruses, underlining the correlation between diagnosed cases in hospitals and viral incidence in viral laboratory testing.

Examining the ICD-10 cases of respiratory viruses across the five hospital departments from October 2015 to December 2023 (Fig. 5b) highlighted seasonal peaks in ED and Outpatient cases, indicating that during RVS, there are consistent spikes in ED and outpatient cases, even before the pandemic. However, post-2020, there was a noticeable shift as ED and outpatient cases remained consistently high, removing any seasonal peaks. In early 2023, a decrease in ED and outpatient cases was observed. However, in mid-2023, there was a spike in ED cases, suggesting a start of a fluctuating trend of hospital visits relating to respiratory viruses.

The ICD-10 annual trend of cases pre- and post-2020 across the hospital departments plot compared respiratory virus cases between 2015-2019 (Fig 6a.) and 2020-2023 (Fig. 6b). In the pre-COVID-19 era, SARS-CoV-2 cases were absent, while RSV and Influenza followed seasonal trends, peaking from November to March. Post-2020, SARS-CoV-2 cases exhibited increasing trends from 2020 to 2022, particularly in ED and outpatient departments, followed by a decrease from 2022 to 2023. Influenza showed a notable spike in ED cases from 2021 to 2022, followed by a decrease from 2022 to 2023. These trends underscore the changing landscape of respiratory virus cases and the varying impacts of COVID-19 on hospital visits across the respiratory viruses.

3.2 Autocorrelation

Autocorrelation is the next crucial step for forecasting in time series analysis. It measures the correlation between observations of a variable at different time points within the time series. It assesses how a variable correlates with itself over time; for example, the correlation of the variable, incidence of RSV, from month one versus month 10 is lag 10. This is particularly important because it helps us understand any temporal patterns or dependencies in the data.

The autocorrelation function (ACF) plot for respiratory virus incidence (Fig. 7) shows the correlation between the respiratory virus's incidence at 100 lags. SARS-CoV-2 shows a high autocorrelation in the initial lags, indicating a strong correlation between adjacent months. A strong positive correlation means high values are followed by other high values, or low values follow low values. However, at lag 30, the autocorrelation crosses zero and into negative autocorrelation, suggesting a change in the underlying trend or behavior of the COVID-19 incidence. The change in correlation direction can highlight a potential shift in trends.

On the other hand, RSV and Influenza show oscillating curves above and below the zero line. This oscillating behavior indicates a cyclical or seasonal pattern in the autocorrelation for both respiratory viruses. Understanding this pattern is crucial and can aid in developing a more accurate forecasting model that includes seasonal variations.

The autocorrelation function plot for the ICD-10 cases looked at the autocorrelation of each specific virus type (Fig. 8a) and hospital departments (Fig. 8b). For hospital departments, ICU and inpatient department ICD-10 cases exhibit similar oscillations in positive autocorrelations at lower lag values but shift to negative autocorrelation around lag 30. Death ICD-10 cases show a decreasing linear trend in positive autocorrelation, eventually crossing the zero line at lag 30 and continuing in negative autocorrelation. ED and outpatient both show oscillating patterns over the zero-line, indicating a cyclical pattern that might correlate with the RVS peaks.

For the specific viruses (Fig 8a), patterns such as RSV, Influenza, and unspecified viral cases displayed similar oscillations in positive autocorrelation initially but shifted to negative autocorrelation at higher lags. SARS-CoV-2 cases also shift from positive to negative autocorrelation, indicating a change in its temporal dynamics. The ACF curve dies off at 40 lags, most likely due to SARS-CoV-2 having fewer months of data than the other respiratory viruses.

Overall, these autocorrelation analyses provide valuable insight into cyclical patterns and any shifts in the data. Understanding these patterns is an essential step for developing forecasting models that capture the complex seasonality of each virus over time.

3.3 Simple Naïve and Seasonal Naïve Forecasting

3.3.1 Incidence Naïve Forecasting

Simple Naive Forecasting (NF) is a straightforward time series forecasting technique where the variable's future value is predicted to equal the last observed value. It assumes the series will continue in its current trend without any major adjustments or abnormalities. On the other hand, Seasonal NF extends this forecasting concept by considering seasonal patterns. It predicts the future value to be equal to the last observed value from the season in the previous year.

The Simple NF model for the incidence of respiratory viruses (Fig. 9) displays the overall incidence from October 2015 to December 2023 with the average and Simple NF for each respiratory virus on the right: RSV, Influenza, and SARS-CoV-2. The lighter of the two colors is the Simple NF, and the darker is the average. The Simple NF for RSV is relatively low compared to the overall incidence and other viruses. The NF for SARS-CoV-2 and Influenza are notably higher most likely because the model takes the last observed value and replicates it for the prediction values.

The Seasonal NF model for the same incidence data (Fig. 10) shows the overall Seasonal NF with each respiratory virus's Seasonal NF individually. The Seasonal NF for RSV and

Influenza are relatively aligned with the overall incidence trend. However, the forecast for SARS-CoV-2 is notably higher than the overall incidence and other virus Seasonal Naïve Forecasts.

3.3.2 ICD-10 Cases Naïve Forecasting

The Simple NF model for ICD-10 cases shows both the respiratory viruses (Fig. 11a) and hospital departments (Fig. 11b). The overall ICD-10 cases are from October 2015 to December 2023 and the Simple NF for respiratory viruses forecast for unspecified viral, RSV, Influenza, and SARS-CoV-2 are on the right. Both the Simple NF and average lines for SARS-CoV-2 ICD-10 cases are notably higher. The other viruses are close together and similar to the overall ICD-10 case forecast. The Simple NF for hospital departments (Fig. 11b) shows that the Simple NF for ED and Outpatient are higher than the other departments.

The Seasonal NF model for ICD-10 cases for respiratory viruses depicts the overall ICD-10 case Seasonal NF (Fig. 12) and explores each virus's forecast individually in separate plots. The Seasonal NF for SARS-CoV-2 is notably higher than the forecasts for other viruses. The Seasonal NF model for ICD-10 cases by departments (Fig. 13) showcases notably higher ICU and outpatient Seasonal NF than other departments. Conversely, the other departments, including inpatient, death, and ED, showed a similar range for their Seasonal NF, indicating a stable pattern in ICD-10 cases across those department categories.

3.4 ARIMA Forecasting

The ARIMA model is a time series forecasting technique that combines autoregression (AR), differencing (I), and moving averages (MA). It is designed to explore trends in time series data and is effective when data is non-stationary, where the mean and variance change over time.

In the context of this analysis, the original data was transformed into logarithmic-data, and all the ARIMA forecast models were conducted on log-incidence or log-cases. This helps show fluctuations and trends for datasets that have large variations in scales. Without log-transformations, the prediction models will have a flattening effect.

3.4.1 Incidence ARIMA Forecasting

The ARIMA prediction model for incidence by respiratory viruses (Fig. 14) shows the overall incidence ARIMA forecast with each respiratory virus: RSV, Influenza, and SARS-CoV-2. Each forecast includes a confidence interval ribbon for the forecast, providing a range of possible outcomes. The ARIMA forecast for the overall incidence is higher than the individual virus forecasts. However, the wide confidence interval indicates considerable uncertainty in the overall forecast, allowing it to include the forecasts of the individual viruses. The viruses' forecasts are relatively close and fall within their respective confidence intervals.

3.4.2 ICD-10 Cases ARIMA Forecasting

The ARIMA prediction model for overall ICD-10 cases for respiratory viruses (Fig. 15) shows the overall ICD-10 case forecast is lower than the individual virus forecast, but the wide

confidence interval ribbon encompasses all the other virus forecasts. Notably, the forecast for SARS-CoV-2 ICD-10 cases is the highest in the viruses, aligning with Figure 5a and the rising trend of COVID-19 cases by the end of 2023.

The hospital department ICD-10 cases ARIMA predictions (Fig. 16) have the same layout as the other ARIMA prediction models. The ARIMA forecast of the overall ICD-10 cases is lower than the ICD-10 department cases' forecasts. However, the wide confidence interval allows the overall forecast to include only the inpatient, ICU, and death ICD-10 cases in the upper bounds of the confidence interval ribbon. The forecasts for outpatient and ED cases have higher forecasts, aligning with the high trends observed in ED and outpatient from 2021 to 2023 in Figure 5b.

3.5 SARIMA Forecasting

The SARIMA prediction model is the Seasonal Autoregressive Integrated Moving Average, an extension of the ARIMA model that includes a fourth parameter to capture seasonality in the data. It accounts for both non-seasonal and seasonal aspects of the data.

3.5.1 Incidence SARIMA Forecasting

The SARIMA prediction model for respiratory virus incidence (Fig. 17) depicts the overall incidence forecast with the surrounding plots of the specific respiratory viruses: RSV, Influenza, and SARS-CoV-2. The training data spans from October 2015 to December 2023. Each SARIMA forecast replicates the shape of a repetitive sinusoidal function, which spans from 2023 to 2032 and includes a confidence interval ribbon. The overall incidence SARIMA forecast shows a higher

amplitude for each period, indicating a wider range of possibilities. Notably, RSV and Influenza forecasts exhibit very similar periods and amplitude heights, suggesting comparable seasonal patterns. On the other hand, the SARIMA forecast for SARS-CoV-2 shows random fluctuations but follows a more linear model rather than a sinusoidal function curve.

3.5.2 ICD-10 Cases SARIMA Forecasting

In Figure 18, the SARIMA prediction model focuses on ICD-10 cases for respiratory viruses. Notably, the overall ICD-10 cases forecast exhibits a larger amplitude height and a wider range for the confidence interval. However, the confidence interval peaks only include the Influenza SARIMA forecast. The SARIMA forecast for SARS-CoV-2 has the highest forecast and narrowest amplitude and confidence interval.

The SARIMA prediction model for ICD-10 cases across the hospital departments has the same format as the two previous figures (Fig. 19). The overall ICD-10 cases forecast exhibits the largest amplitude height and broader range for the confidence interval. However, the confidence interval only partially includes the SARIMA forecast for death ICD-10 cases. Inpatient, outpatient, and ED forecasts share similar sinusoidal curve shapes, amplitudes, and periods, indicating comparable patterns across these departments.

3.6 Prediction Model Performance Metrics

Prediction model performance metrics play an essential role in assessing the efficacy and accuracy of different forecasting models. Although we knew about the seasonal trend in the RVS,

we tested several models: Simple NF, Seasonal NF, ARIMA, and SARIMA. The performance metrics used to evaluate the performance of each model include AIC (Akaike Information Criterion), RMSE (Root Mean Squared Error), MSE (Mean Squared Error), and MAE (Mean Absolute Error).

AIC, RMSE, MSE, and MAE all indicate a better performing model with lower values. Each table shows the performance metrics for each prediction model. Each table looks at the datasets: the incidence of respiratory viruses (Table 1), ICD-10 cases by hospital departments (Table 2), and ICD-10 respiratory virus cases (Table 3).

The results consistently showed that SARIMA outperformed the other models for all three datasets based on the four performance metrics. SARIMA's ability to use both non-seasonal and seasonal parameters made it the most suitable for both datasets. Interestingly, in a few instances, the Seasonal NF exhibited better performance metrics than the ARIMA model.

4.0 Discussion

Our study used predictive modeling to understand the dynamics of pandemics and predict how COVID-19 interacts with the respiratory virus season. The 1918 influenza pandemic and the SARS-CoV-2 pandemic share parallels in their global impact, high contagiousness, and strain on the healthcare systems; however, a significant difference was the affected demographic [6]. The lessons that can be integrated from the 1918 pandemic into resource allocation for future potential pandemics are early detection and response strategies, like vaccines, and strengthening healthcare[6]. The purpose of this study is to use predictive modeling to understand the impact COVID-19 had on laboratories and hospitals for future planning and resource allocation.

While our analysis indicates a notable shift in normal RVS dynamics coinciding with the emergence of COVID-19, it is crucial to acknowledge that other factors may have contributed to these observed changes. Our predictive models provide insights into trends but do not definitively establish a causal relationship between COVID-19 and the shift in RVS dynamics. Further research is needed to reveal the specific impact of COVID-19 and other potential factors on the observed shifts in RVS dynamics.

Examining respiratory virus incidence and ICD-10 cases from pre-COVID-19 to post-COVID-19 periods uncovers the evolving burden on laboratory and hospital departments. This analysis, spanning from 2015-2019, through the pandemic years of 2020-2021, and into the post-pandemic era from 2022 onward, provides critical insights into managing current healthcare challenges and preparing for potential future outbreaks. For our predictive modeling framework, we used 2022 as the “new normal” post-pandemic data because this was when UPMC lifted its hospital-wide mask mandates.

Exploratory data analysis (EDA) forms the foundational step in our predictive modeling approach, aiding in understanding data structure, identifying patterns, and initiating hypotheses. The EDA of respiratory virus incidence (Fig. 3a) mapped out RSV and Influenza exhibiting seasonal peaks aligning with the typical respiratory viral seasons. The onset of SARS-CoV-2 in 2020 coincided with the disruption of the typical viral seasonal patterns when there was an absence of RSV and Influenza incidence from March 2020 to August 2021. Several factors may explain this phenomenon, such as the transition from RVP tests to 4plex molecular tests, altered practices due to quarantining, a subdued flu season, or the potential dominance of SARS-CoV-2 over the other viruses. The subsequent figure (Fig. 4) delves deeper into seasonal and annual incidence trends and pre- and post-COVID-19 comparisons. These highlight the shift in typical dynamics of the RVS that occurred at the onset of SARS-CoV-2.

The EDA of ICD-10 cases (Fig. 5a and 5b) of respiratory viruses (Fig. 5a) across hospital departments (Fig. 5b) not only reflects the burden of RVS on healthcare systems but also provides insights for resource allocation during pandemics. The EDA for ICD-10 cases of respiratory viruses reveals the short-lived presence of unspecified viral cases only in 2020. This may be caused because physicians initially categorized many respiratory viral infections as unspecified respiratory viruses due to the diagnostic challenges in identifying SARS-CoV-2 through ICD-10 codes. By the start of 2021, these unspecified viral cases diminished, indicating an improved diagnosis of respiratory viruses.

Furthermore, analyzing ICD-10 cases across hospital departments showcases the peaks in ED and outpatient during RVS peaks, but post-2020, ED and outpatient had consistently high ICD-10 cases. The peaks observed in ED and outpatient cases during respiratory virus season (RVS) highlight the increased demand for medical attention during these periods. Allocating additional

resources to these departments during RVS can improve healthcare delivery and relieve strain in healthcare departments.

Exploratory data analysis (EDA) of the incidence and ICD-10 datasets is crucial in the data prediction process because it helps us understand the data structure and identify patterns or abnormalities. Before starting the forecasting analysis, EDA helps us assess the data and initiate any potential hypotheses of the analysis.

Our primary forecasting analysis examined various prediction models, including Simple and Seasonal Naive forecasting, ARIMA, and SARIMA. Simple and Seasonal NF models differ primarily in terms of seasonal variations. Simple NF uses average and naïve future values, which averages the last year's values and recreates that trend as the Simple NF model, assuming a constant trend. However, Seasonal NF considers the cyclical patterns by aligning the predictions with the past seasonal patterns. The incidence's Simple NF show that Influenza and SARS-CoV-2 NF are higher than the overall incidence. This correlates with what appears in Figure 3 because, at the end of 2023, Influenza and SARS-CoV-2 incidence numbers were relatively similar and were rising in incidence. Simple NF averages the last year's values and replicates them as the forecast. The Simple NFs for ED and Outpatient ICD-10 cases (Fig. 11b) were higher than the other departments. This correlates to Figure 5b because, from 2020 onward, ED and outpatient have been consistently high instead of at their typical seasonal peaks. The Simple NF used the last observed 2023 data and recreated those ICD-10 case numbers for ED and outpatient forecasts.

The incidence and ICD-10 Seasonal NF models capture the typical seasonal peaks for RSV and Influenza reasonably well; however, NF struggles to predict the shifting nature and abnormalities of SARS-CoV-2 (Fig. 10 and Fig. 12). Simple and Seasonal NF, while applicable for quick assessments, are generally not ideal models for reliable prediction in time series analysis.

However, like Simple NF, Seasonal NF cannot adapt to sudden changes or shifts in the underlying trend of the data. These models assume the past patterns will remain the same, which is not ideal for the complexities of viral incidence and ICD-10 cases over the years.

All ARIMA and SARIMA prediction models were transformed into log scales. Reversing the logarithmic transformation of log-incidence or log-cases would mean exponentiating the log values, which would flatten the prediction due to the large variation in incidence or case values pre- and post-COVID-19.

ARIMA differs significantly from Simple and Seasonal Naïve Forecasting methods. Simple NF relies only on the most recent observed values to make predictions. In contrast, ARIMA considers the entire retrospective data and incorporates autoregression and moving average components to capture complex patterns. On the other hand, Seasonal NF focuses on seasonal patterns but does not account for irregularities present in the data, which ARIMA can capture. However, some metrics in the performance metrics tables showed that Seasonal NF performed better than ARIMA, even though ARIMA is the more complex model. This could be attributed to the ARIMA model's struggle to capture complex season patterns because it assumes a linear relationship. Conversely, the seasonal NF leverages the seasonal nature by using the last observed value from the season in the previous year as its forecasting training data. While ARIMA is a powerful and widely used forecasting model, there may be better options for time series with strong seasonality patterns. ARIMA models assume linear relationships and may struggle to capture complex season patterns.

SARIMA is an extension of ARIMA prediction models, but it includes a fourth parameter of seasonality. The SARIMA models are displayed with a repetitive sinusoidal shape for the forecasts. It allows for better predictions by including both non-seasonal and seasonal parameters

to explore the complex dynamics of the Respiratory Viral Season. The sinusoidal curve was repetitive for the SARIMA prediction models for each period. Despite SARIMA outperforming other models in the performance metrics, we faced challenges due to the limited post-pandemic data available. We considered the post-pandemic years from 2022-2023. SARIMA models typically require multiple seasonal cycles (4-6 cycles) to capture seasonality effectively, and our data post-COVID-19 only spans 1.5 cycles (2022-2023) [8].

The fluctuating nature of COVID-19, including the emergence of new variants, vaccination and treatments, and changes in public health measures, adds complexity to forecasting efforts. These dynamic factors can significantly impact virus transmission rates, severity of cases, and healthcare system burden, making accurate long-term predictions challenging. Therefore, our findings emphasize the need for continued data collection to differentiate whether COVID-19 is transitioning into an endemic virus similar to Influenza. While these predictive models offer valuable insights into visualizing trends, they also have limitations, especially for predicting post-COVID-19 pandemic trends.

4.1 Limitations

The QI project encountered several limitations that should be acknowledged. Firstly, we had to make certain assumptions, which may have influenced the accuracy of our findings. We assumed that the molecular viral tests we used were 100% sensitive even though such tests are not infallible, potentially leading to both false positives and false negatives. Additionally, we assumed 100% accuracy in ICD-10 codes as diagnostic indicators to measure the hospital's burden, overlooking the inherent variability in coding practices among healthcare professionals. Different

physicians may favor specific codes for diagnosis, introducing potential inconsistencies in the data. Our analysis revealed disparities in testing practices over time, with under-testing of respiratory viruses with RVP tests before the COVID-19 pandemic and potential over-testing after the pandemic using the 4plex Molecular tests [10]. Pre-pandemic, testing for respiratory viruses like influenza was typically reserved for severe cases or high-risk individuals, such as the elderly or immunocompromised, while routine testing of mild symptoms was less common [11]. These limitations highlight the need for cautious interpretations of the results and the importance of addressing these challenges in future research endeavors.

4.2 Public Health Implications

The insights gained from our QI project hold significant implications for understanding the burden of viral respiratory infections on healthcare facilities. Our future direction includes collecting more post-COVID-19 RVS data to develop an effective predictive tool designed to predict the healthcare and laboratory burdens associated with respiratory virus seasons. This tool will utilize the insights gained from our data to predict the burden on healthcare systems, enabling proactive resource allocation. Additionally, our research aims to ascertain whether COVID-19 remains an additional burden or potentially becomes an integral part of the regular respiratory season. This is a vital step toward proactive healthcare planning and resource allocation to ensure healthcare facilities' continued resilience and efficiency.

5.0 Acknowledgements

I would like to thank Dr. William Pasculle for his mentorship and assistance in navigating the data collection process and securing permissions from UPMC. I would like to thank Dr. Alan Wells for his support during this project. Additionally, I am grateful for Dr. Joseph Yurko for his expertise in data analytics and programming, which helped me figure out the analytical aspects of this study. Lastly, I would like to thank Dr. Jeremy Martinson for his role as my Master's Essay Chair and for providing valuable feedback through this process.

6.0 Figures and Tables

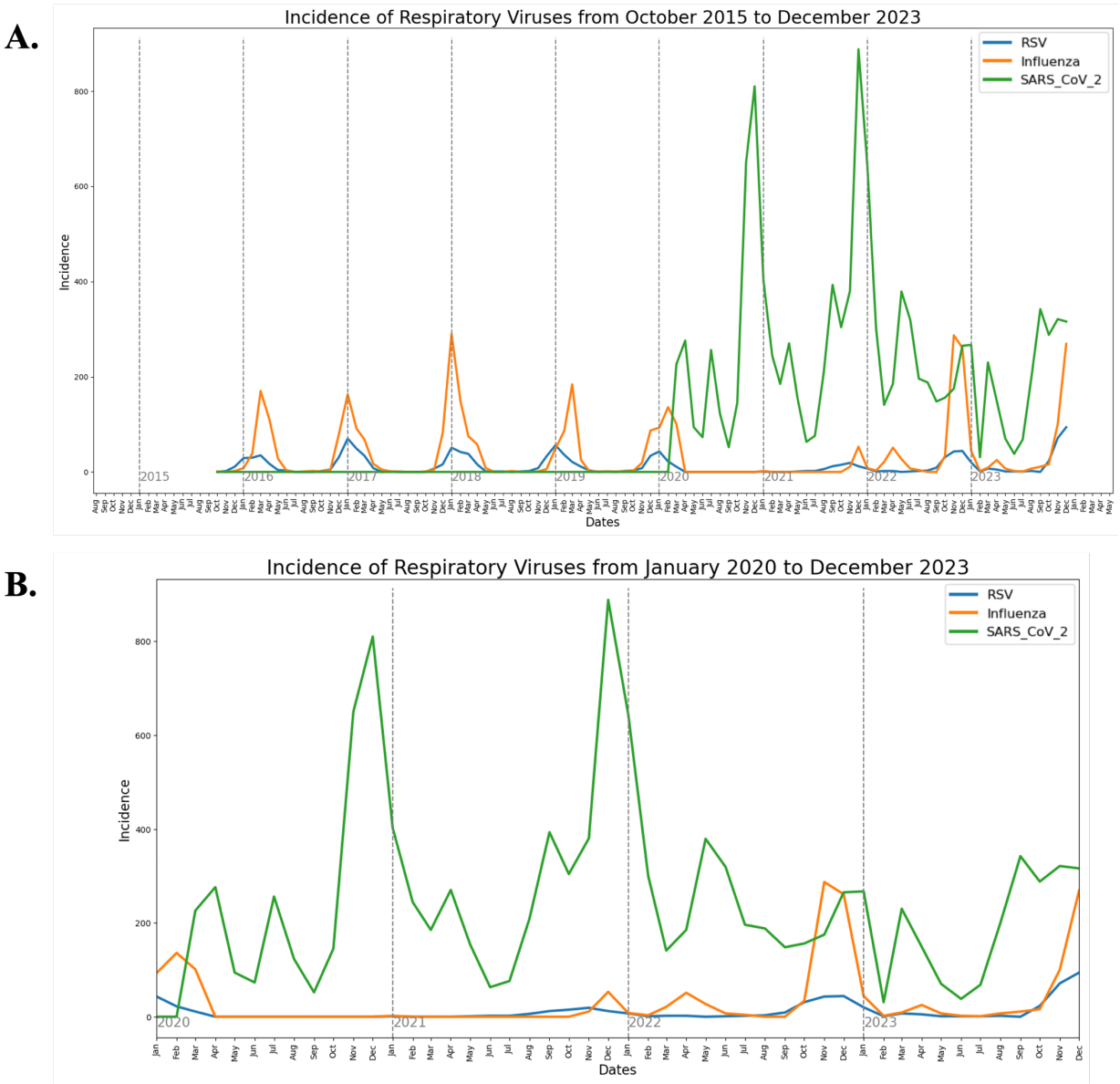


Figure 3. Respiratory Virus Incidence (2015-2023).

Graphs highlighting the incidence of respiratory viruses within a specific timeframe. A. Incidence of Respiratory Virus from October 2015 to December 2023 provides a comprehensive overview of respiratory virus incidence spanning from 2015 to 2023 of three respiratory viruses: RSV, Influenza, and SARS-CoV-2. B. Respiratory Virus Incidence 2020-2023: narrows the focus to the period between 2020 and 2023.

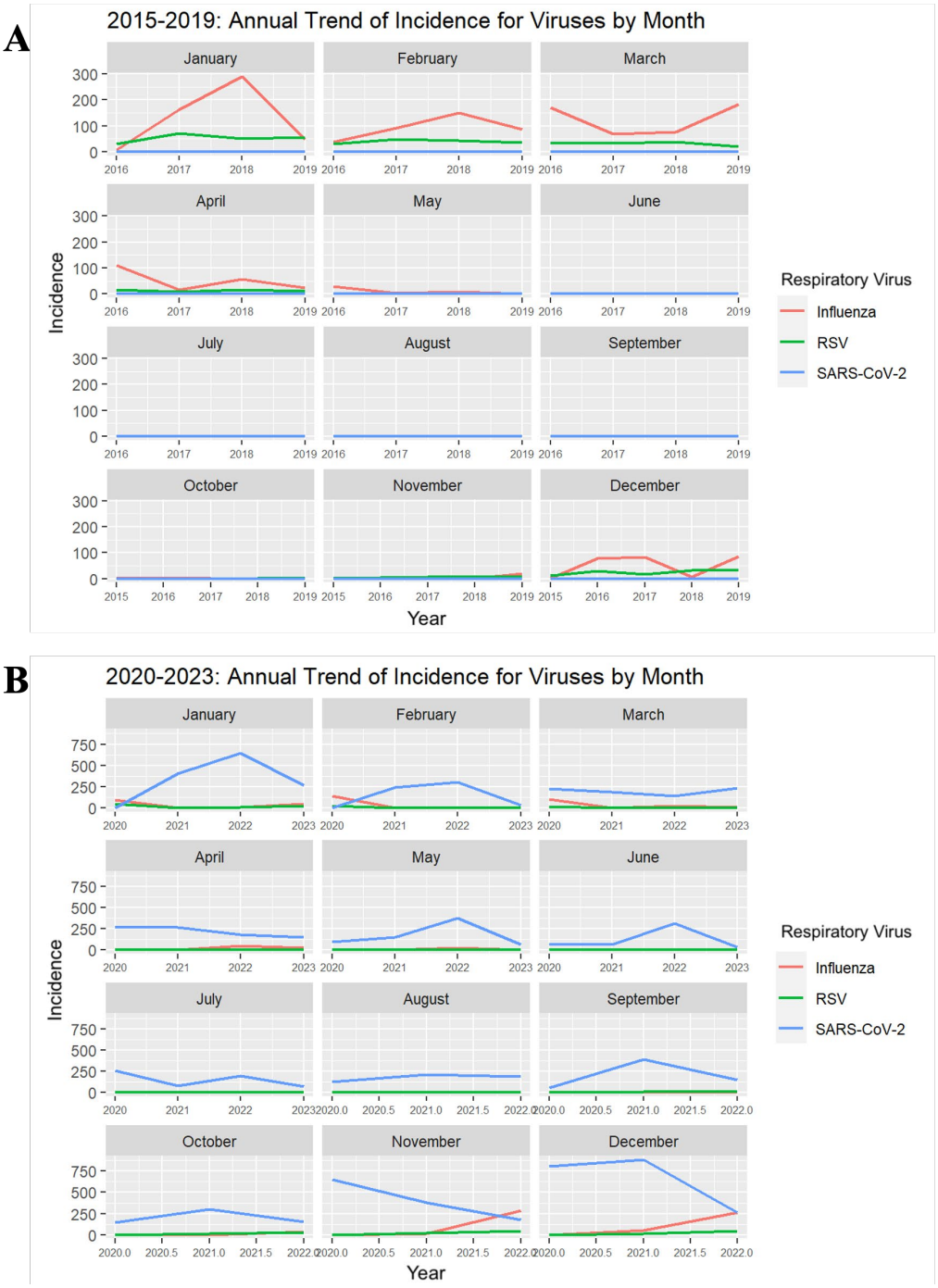
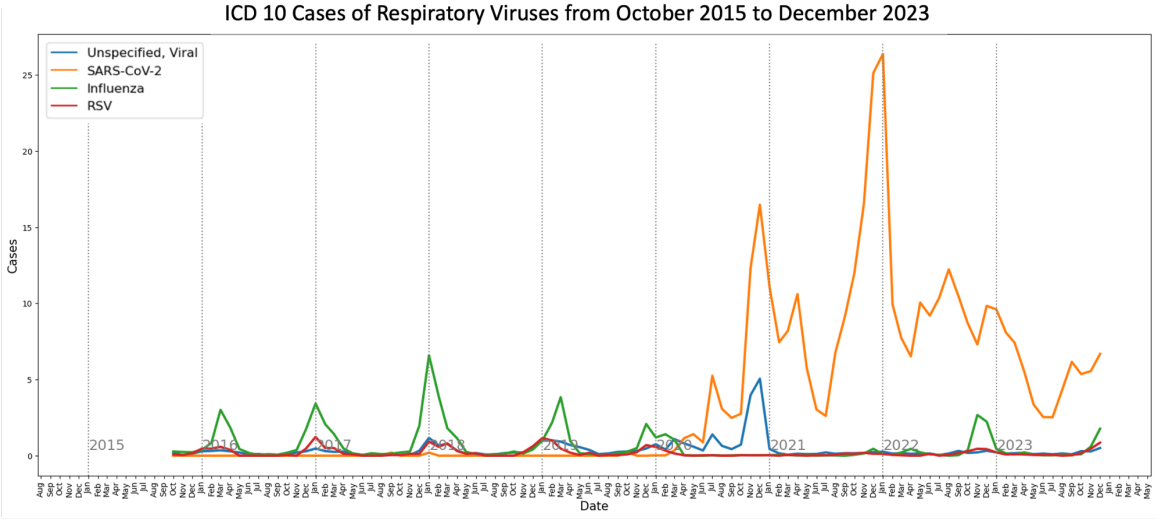


Figure 4. Annual Trend of Incidence Pre- and Post-SARS-CoV-2.

Two time series plots to compare the annual trends in incidence for three respiratory viruses: RSV (green), Influenza (red), and SARS-CoV-2 (blue), over two timeframes: 2015-2019 and 2020-2023. These plots remove the seasonal variations and focus on the trends of each virus's incidence within specific months.

A.



B.

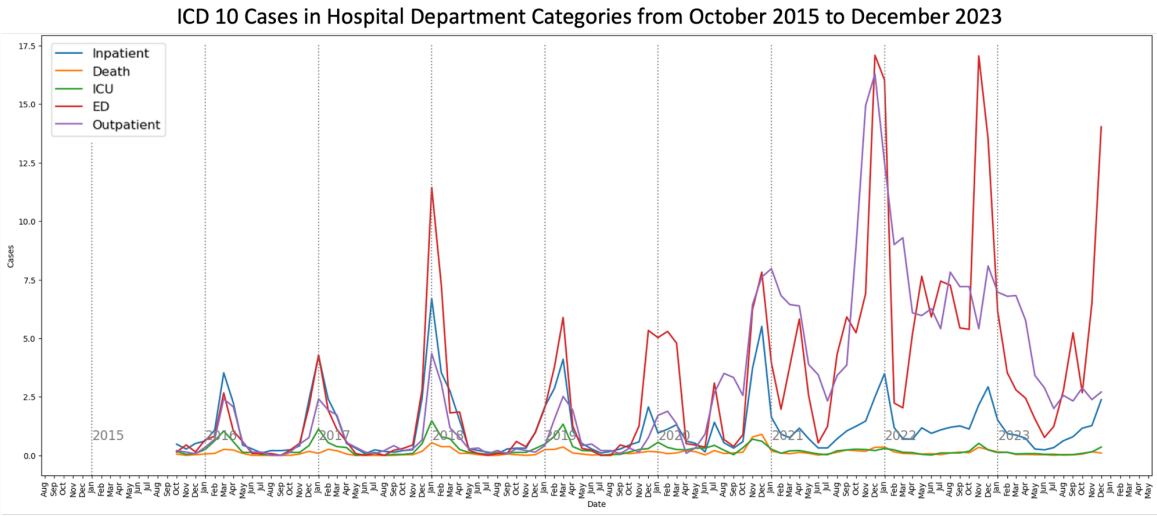


Figure 5. ICD-10 Respiratory Virus Cases from October 2015 to December 2023.

A. Graph that shows the respiratory virus ICD 10 cases across the five different hospital department categories: Inpatient (blue), Death (orange), ICU (green), ED (red), Outpatient (purple). B. Graph that shows the respiratory virus ICD 10 cases divided up in the respiratory viruses: RSV (red), Influenza (green), SARS-CoV-2 (orange), and Unspecified, Viral (blue).

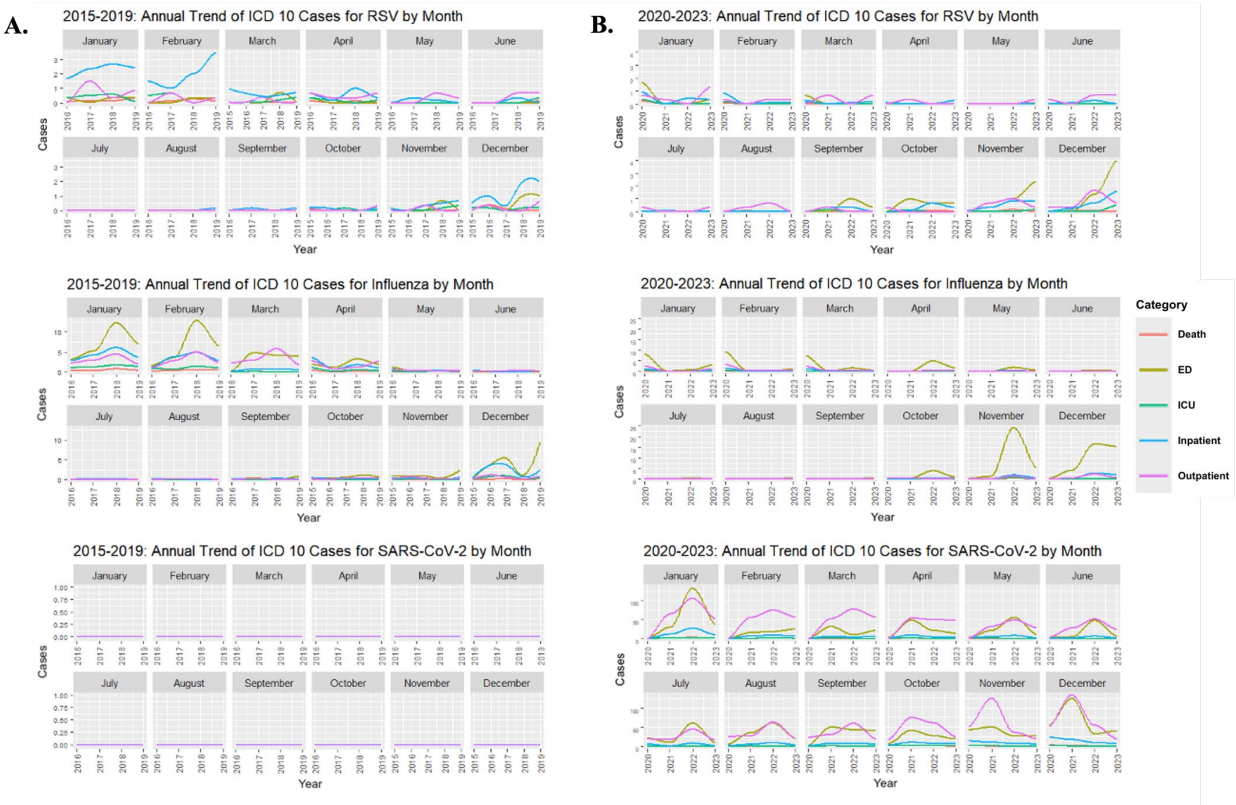


Figure 6. Annual Trend of ICD-10 Cases Pre- and Post-COVID-19 Across Departments.

Time series plots to compare the annual trends in ICD 10 cases for the three respiratory viruses: RSV, Influenza, and SARS-CoV-2 over two timeframes: 2015-2019 and 2020-2023. These plots look at the annual trends across hospital department categories: Death (red), ED (yellow), ICU (green), inpatient (blue), and outpatient (purple). These plots remove the seasonal variations and focus on the trends of each virus's incidence within specific months.

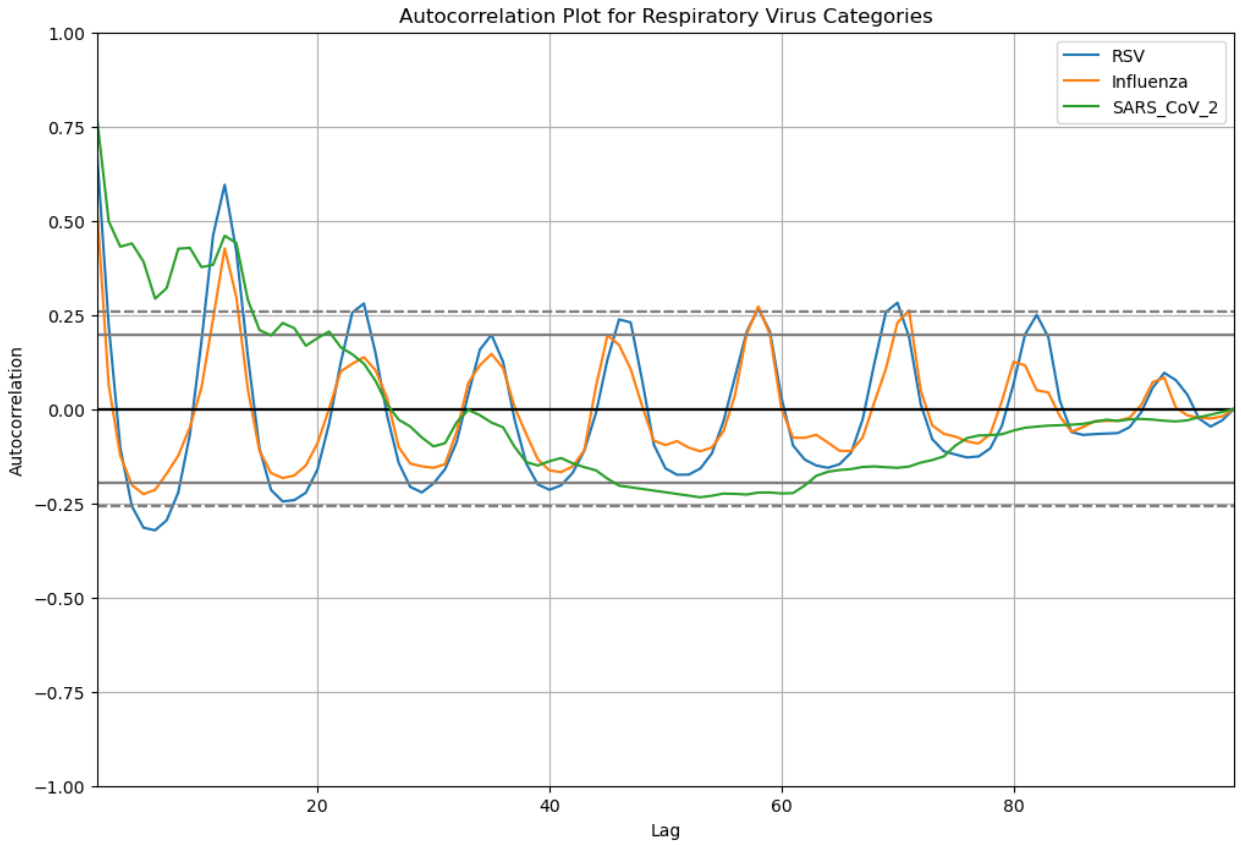


Figure 7. Incidence Autocorrelation Function for Respiratory Virus.

Autocorrelation functions (ACF) of respiratory virus incidence from October 2015 to December 2023, with a lag parameter set to 100. Each curve represents the autocorrelation pattern for a specific virus: Respiratory Syncytial Virus (RSV) is depicted in blue, Influenza in orange, and SARS-CoV-2 in green.

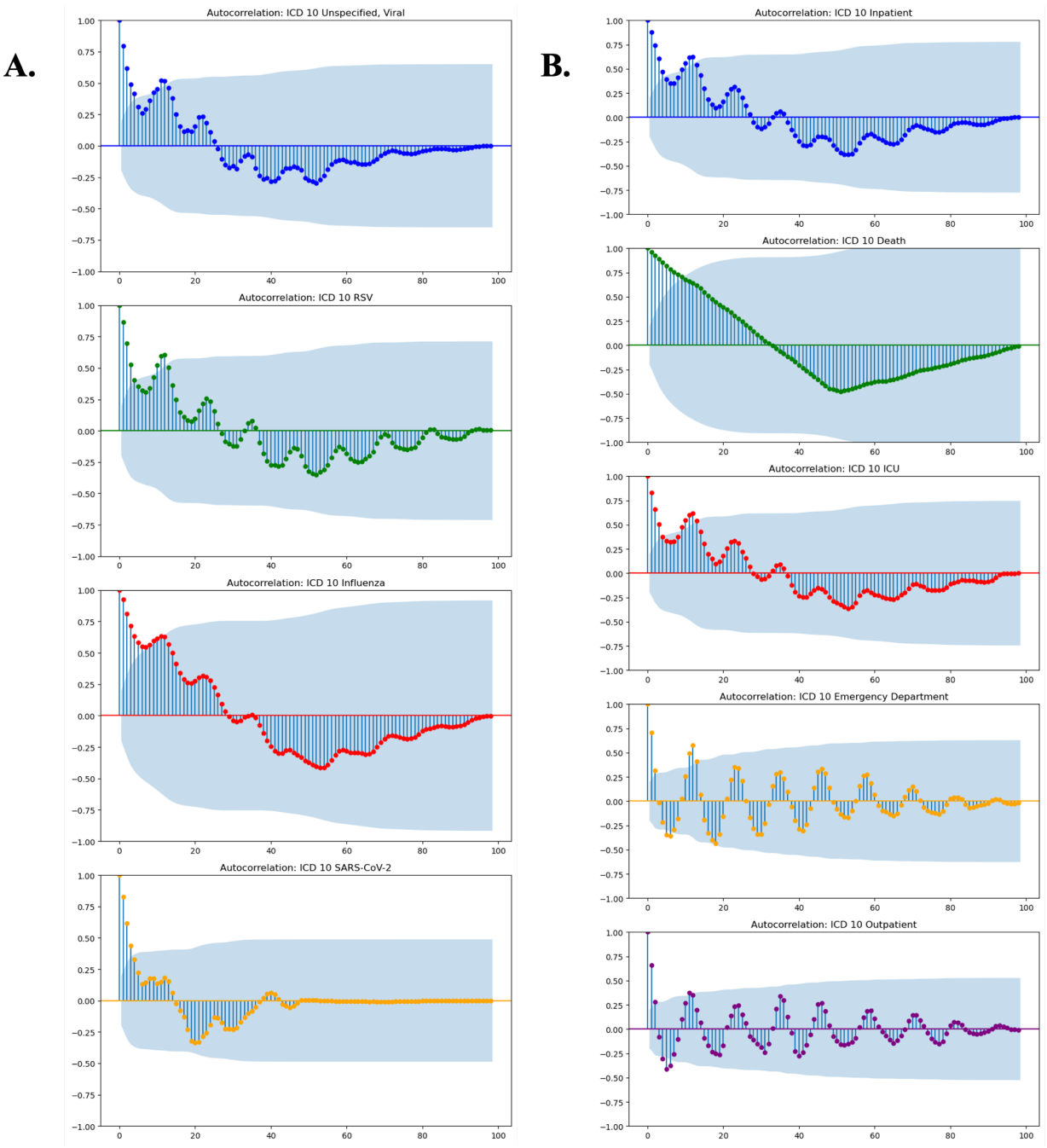


Figure 8. Autocorrelation Function for ICD-10 Cases.

ACF of ICD-10 cases from October 2015 to December 2023, with a lag parameter set to 100. A. ACF of the ICD 10 cases for respiratory viruses. Each ACF represents the autocorrelation pattern for Unspecified, Viral, RSV, Influenza, and SARS-CoV-2, respectively. B. ACF of the ICD 10 cases in hospital department categories. Each ACF represents the autocorrelation pattern for each department category: Inpatient, Death, ICU, ED, and Outpatient, respectively.

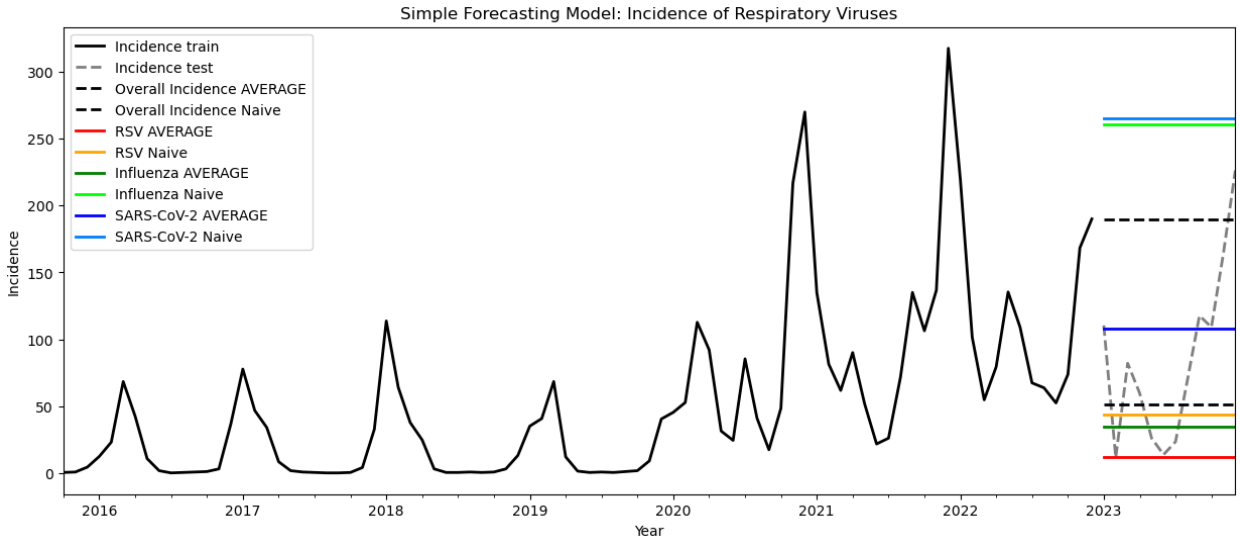


Figure 9. Simple Naive Forecasting Model: Incidence of Respiratory Viruses.

A plot that shows the respiratory virus incidence forecast using a simple naive forecasting model from October 2015 to December 2023. The overall incidence forecast is represented by the black line. The color lines show the average and simple naive forecast for each virus: RSV, Influenza, and SARS-CoV-2.

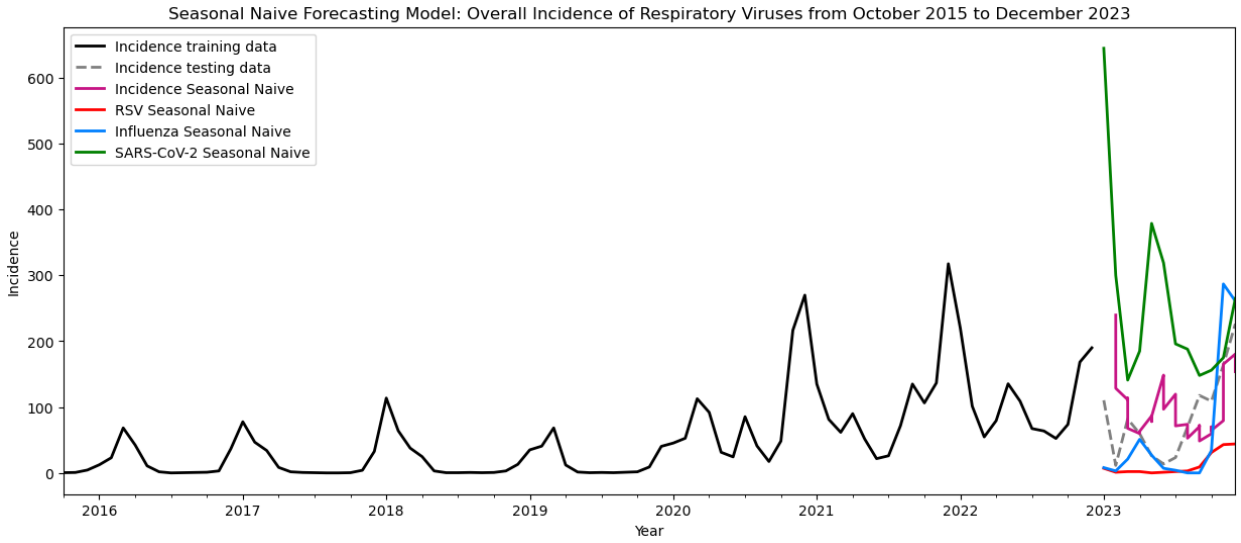


Figure 10. Seasonal Naive Forecasting Model: Incidence of Respiratory Viruses.

The color lines show the seasonal naive forecast for a specific respiratory virus: RSV (red), Influenza (blue), and SARS-CoV-2 (green). The magenta line is the overall seasonal naive forecast for all respiratory viruses. The solid black line is the training data to building the forecasting model, and the dotted line is the testing data used for validation.

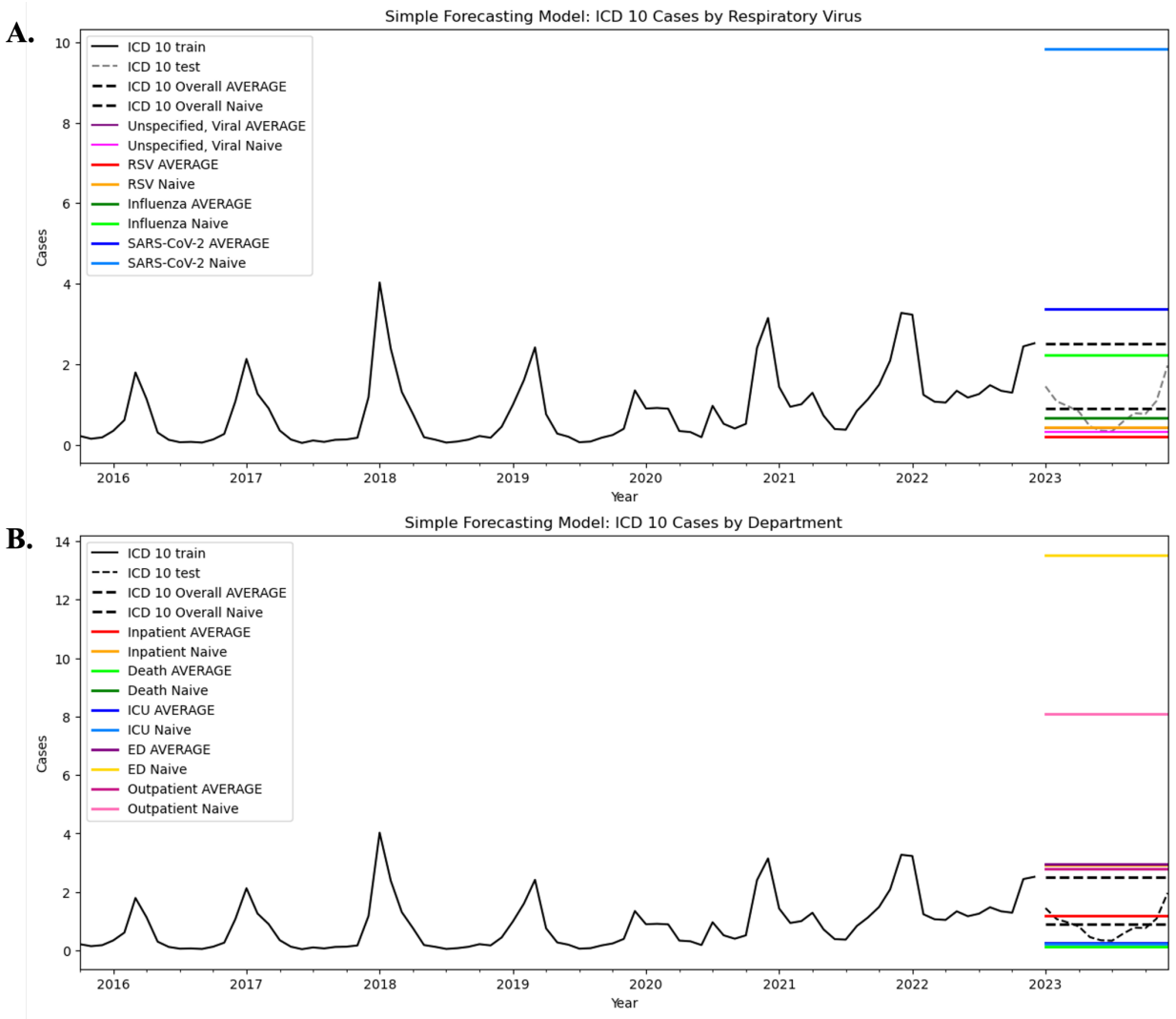


Figure 11. Simple Naive Forecasting Model: ICD-10 Cases.

A.) This plot shows the Simple Naive Forecast for ICD 10 cases by respiratory viruses: Unspecified, Viral, RSV, Influenza, and SARS-CoV-2. The color lines show the average and simple naive forecast for each virus.

B.) This plot shows the Simple Naive Forecast of ICD 10 cases by hospital departments: inpatient, death, ICU, ED, and Outpatient. The color lines show the average and simple naive forecast for each department.

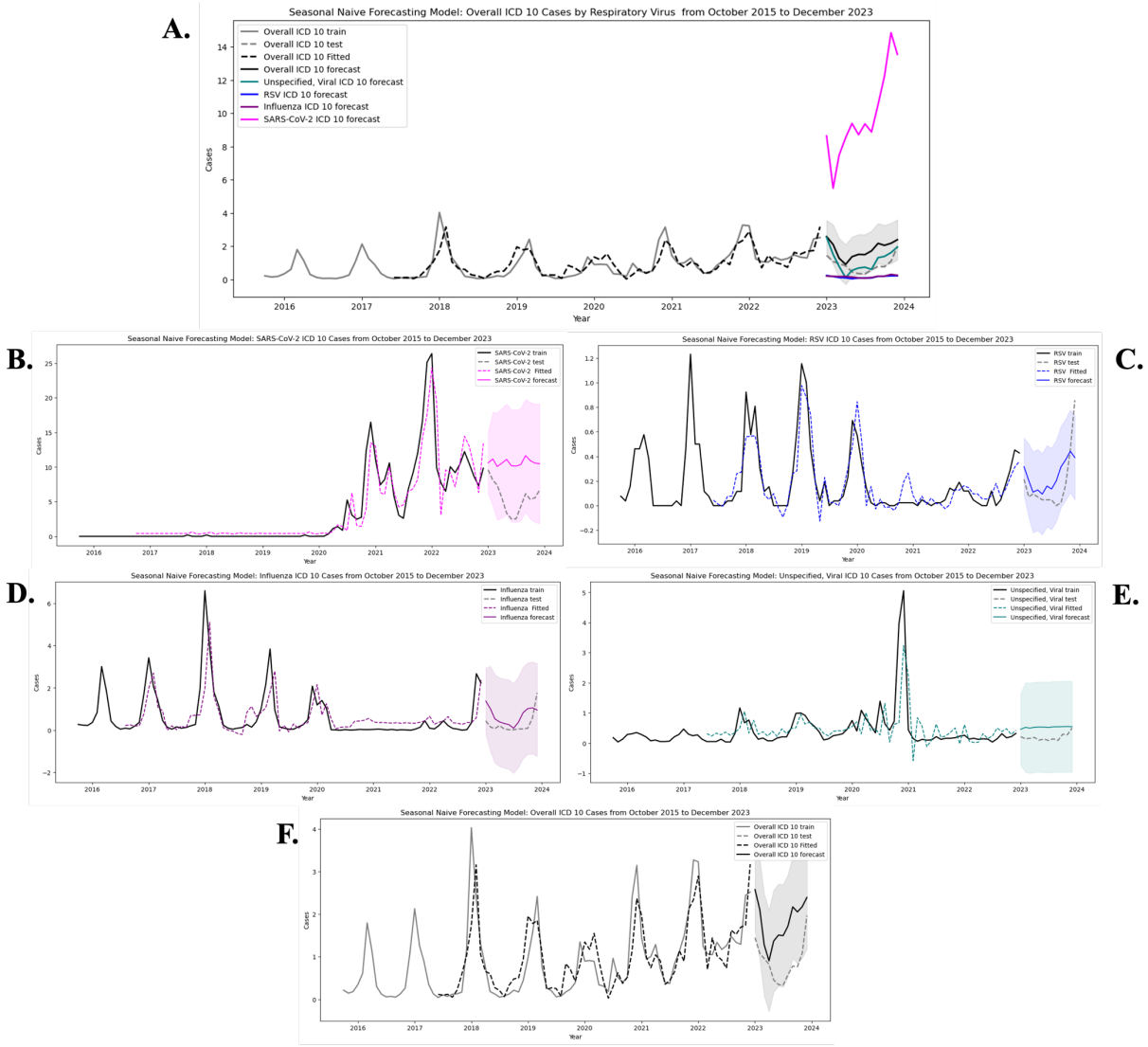


Figure 12. Seasonal Naive Forecasting Model: ICD-10 Cases by Respiratory Viruses.

The color lines show the seasonal naive forecast for a specific respiratory virus: Unspecified, Viral (teal), RSV (blue), Influenza (purple), and SARS-CoV-2 (magenta). A. The plot shows the overall ICD-10 cases seasonal naive forecast, with each respiratory virus forecasts. The solid black line is the training data to build the forecasting model, and the dotted line is the testing data used for validation. The surrounding plots show each respiratory virus and focus on their respective seasonal naive forecast models.

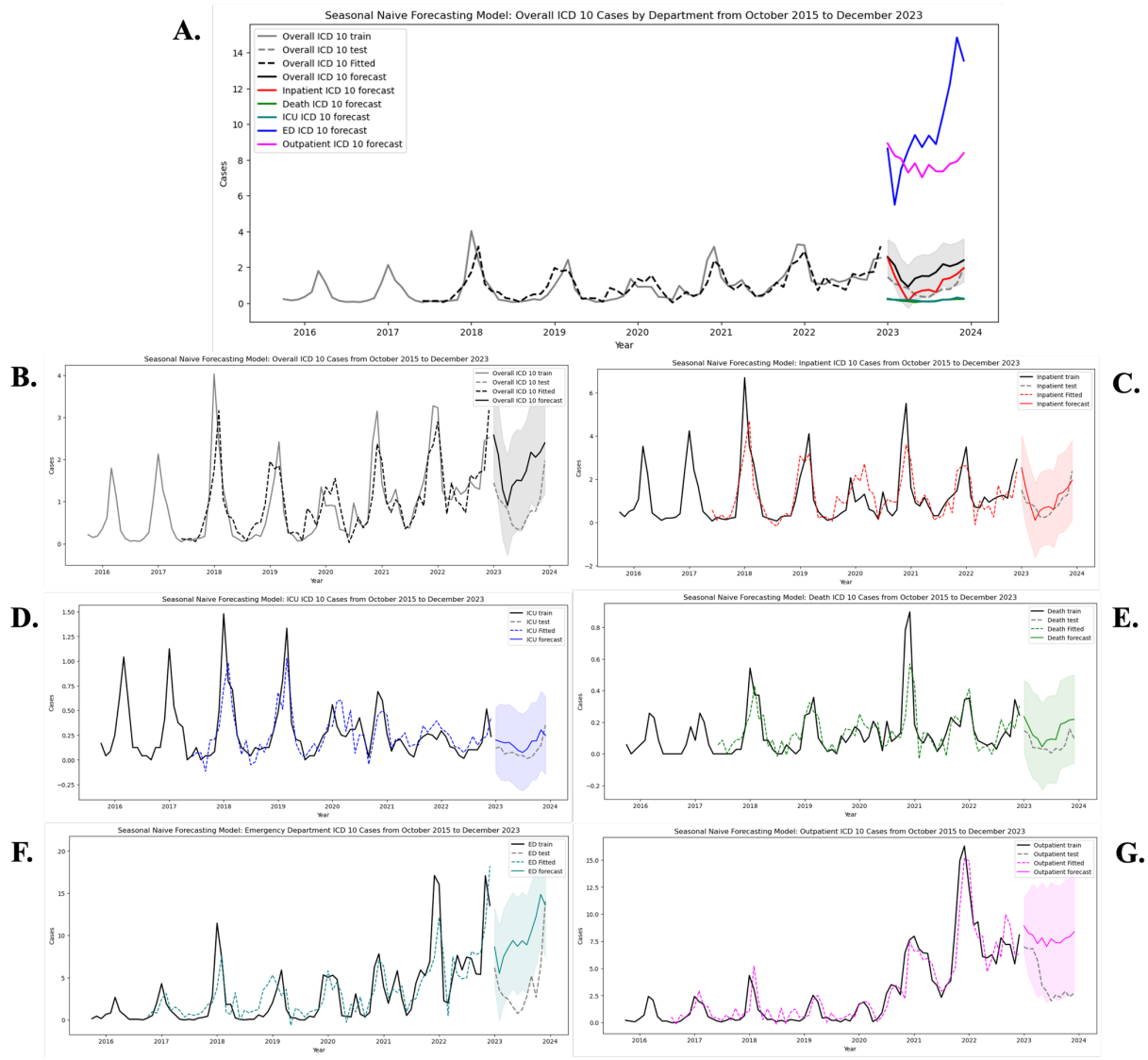


Figure 13. Seasonal Naive Forecasting Model: ICD-10 Cases by Departments.

The color lines show the seasonal naive forecast for a specific department: Inpatient (red), death (green), ICU (blue), ED (teal), outpatient (magenta). A.) The plot shows the overall ICD-10 cases seasonal naive forecast, with each department forecasts. The solid black line is the training data to build the forecasting model, and the dotted line is the testing data used for validation. The surrounding plots show each department and focus on their respective seasonal naive forecast models.

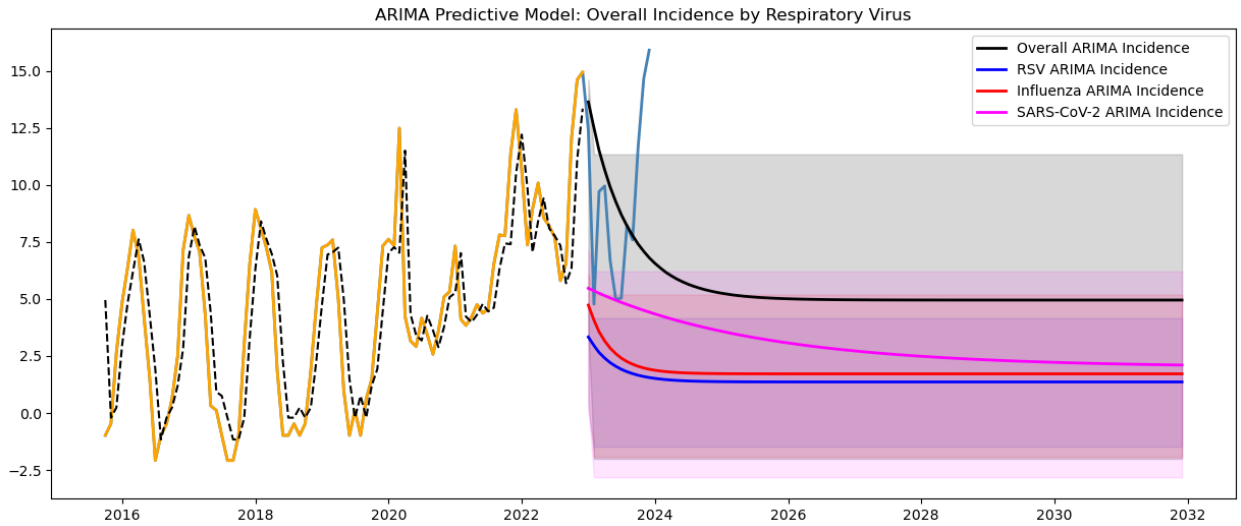


Figure 14. ARIMA Prediction Model for Overall Incidence by Respiratory Viruses.

The incidence ARIMA forecast is from October 2015 to December 2023, with the predictions extending to 2032. The ARIMA forecast of the overall incidence is represented in black and the ARIMA forecasts of all the respiratory viruses: RSV (blue), Influenza (red), and SARS-CoV-2 (magenta), along with its associated confidence interval ribbon in their respective colors. The training data used to train each ARIMA model is represented in orange and spans from October 2015 to December 2023.

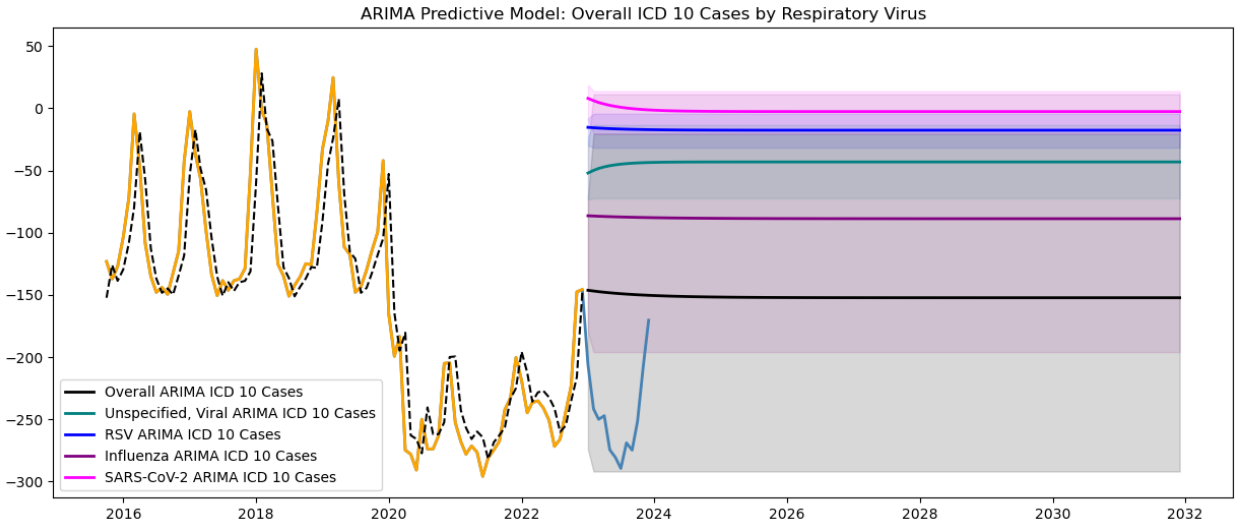


Figure 15. ARIMA Prediction Model for Overall ICD 10 Cases for Respiratory Viruses.

The ICD-10 Respiratory Viruses ARIMA forecast is from October 2015 to December 2023, with the predictions extending to 2032. The ARIMA forecast of the overall ICD 10 cases is represented in black, and the respiratory viruses: Unspecified, Viral (teal), RSV (blue), Influenza (purple), and SARS-CoV-2 (magenta), along with its associated confidence interval ribbon in their respective colors. The training data used to train each ARIMA model is represented in orange and spans from October 2015 to December 2023.

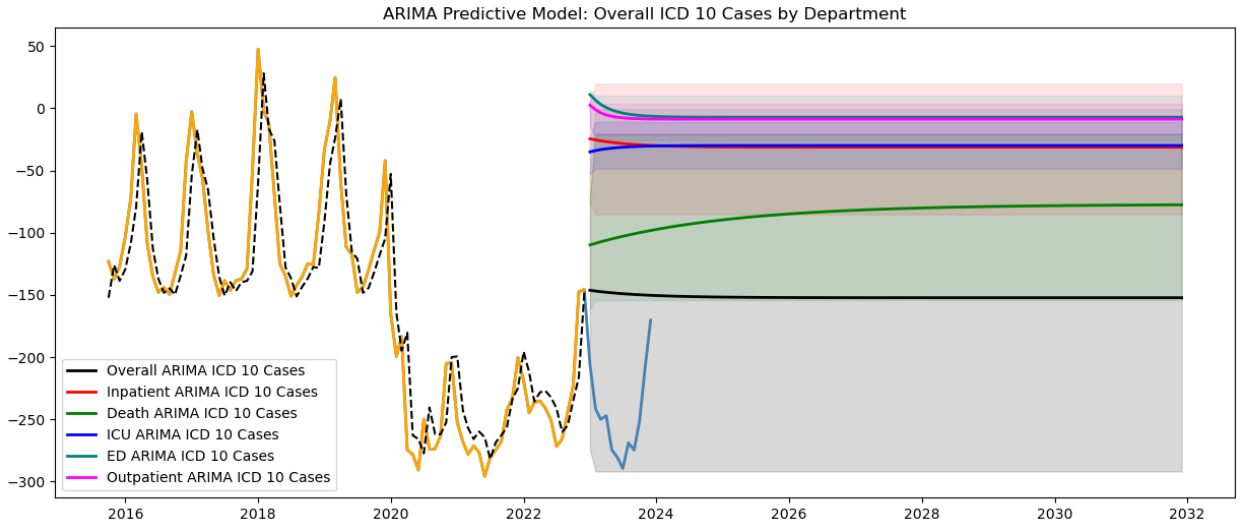


Figure 16. ARIMA Prediction Model for Overall ICD 10 Cases by Hospital Departments.

The ARIMA forecast of ICD-10 Departments is from October 2015 to December 2023, with the predictions extending to 2032. The ARIMA forecast of the overall ICD 10 cases is represented in black, and the ARIMA forecasts of all the department categories: Inpatient (red), death (green), ICU (blue), ED (teal), and Outpatient (magenta), along with its associated confidence interval ribbon in their respective colors. The training data used to train each ARIMA model is represented in orange and spans from October 2015 to December 2023.

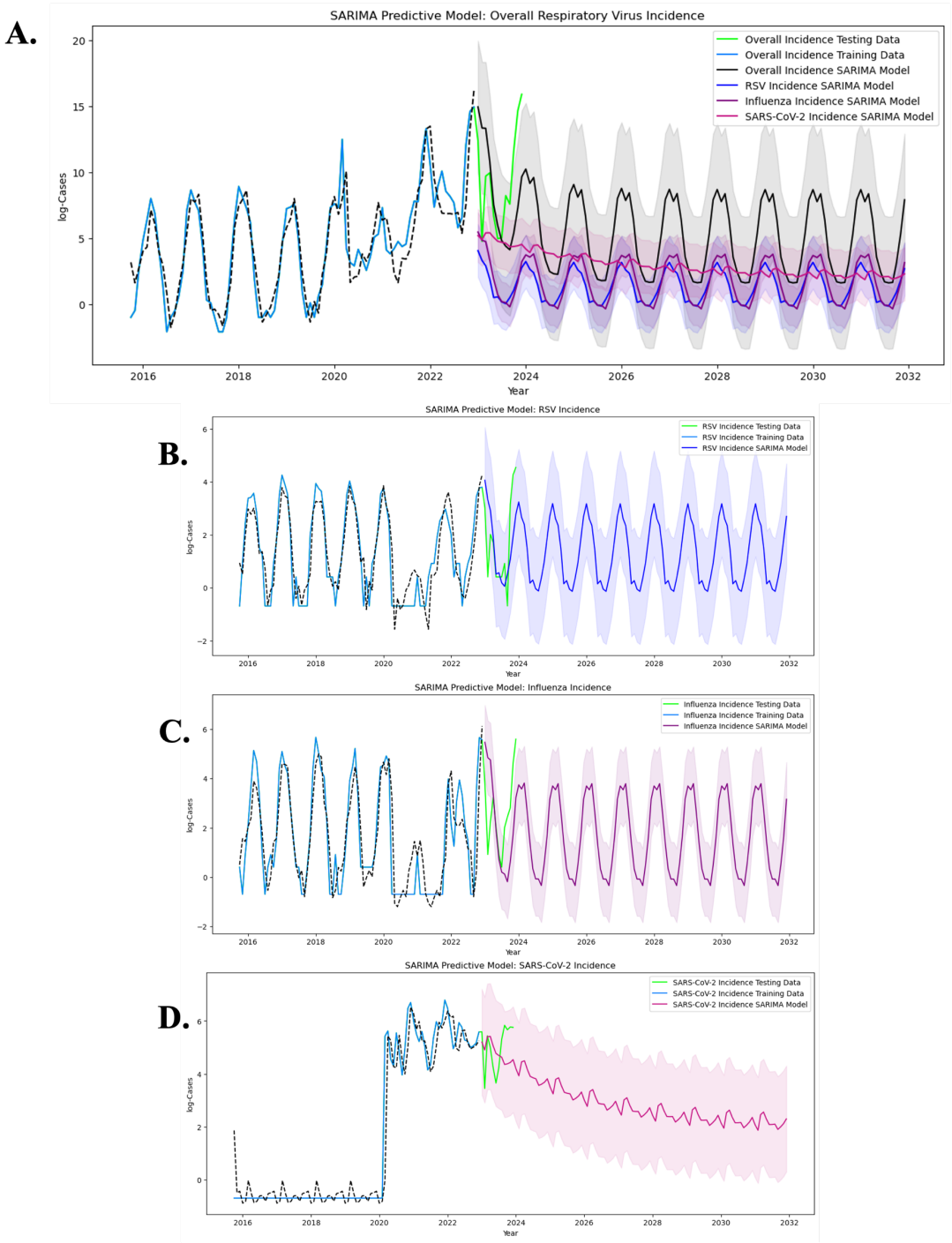


Figure 17. SARIMA Prediction Model for Overall Incidence by Respiratory Viruses.

The main plot shows the SARIMA forecast of the overall incidence with the SARIMA forecasts of all the respiratory viruses. The surrounding plots focused on each respiratory virus: Overall Incidence (black), RSV (blue), Influenza (purple), and SARS-CoV-2 (magenta), along with its associated confidence interval ribbon in their respective colors.

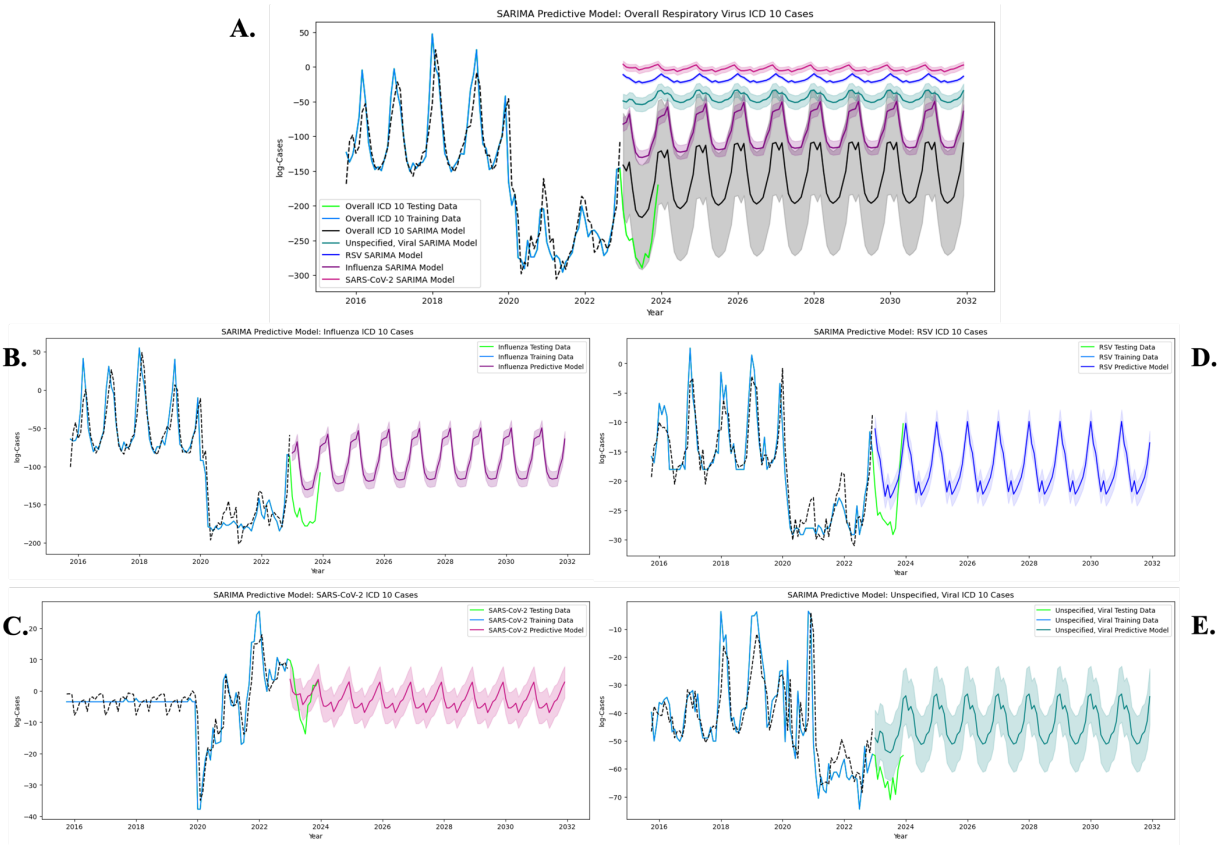


Figure 18. SARIMA Prediction Model for Overall ICD 10 Cases for Respiratory Viruses.

The main plot shows the SARIMA forecast of the overall ICD 10 cases with the SARIMA forecasts of all the respiratory viruses. The surrounding plots focused on each respiratory virus: Overall Incidence (black), RSV (blue), Influenza (purple), and SARS-CoV-2 (magenta), along with its associated confidence interval ribbon in their respective colors.

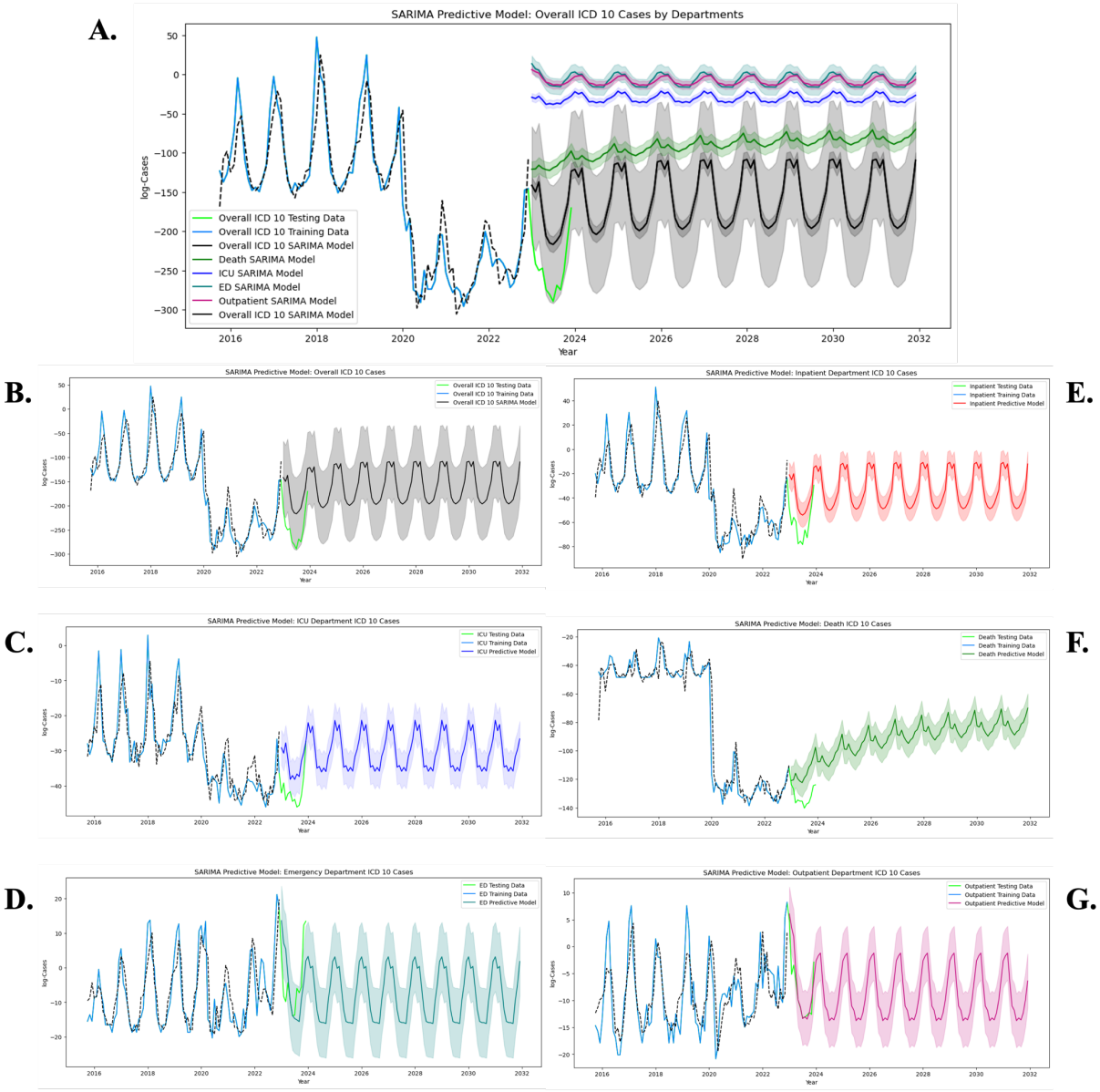


Figure 19. SARIMA Prediction Model for Overall ICD 10 Cases by Departments.

The main plot shows the SARIMA forecast of the overall ICD 10 cases with the SARIMA forecasts of all the department categories. The surrounding plots focused on each department : Inpatient (red), death (green), ICU (blue), ED (teal), and Outpatient (magenta), along with its associated confidence interval ribbon in their respective colors.

Table 1. Performance Metrics for the Incidence Prediction Models

Table 1: Prediction Model Accuracy - Incidence Respiratory Viruses	AIC	RMSE	MSE	MAE
Simple Naive	985.679	436.99	194.32	177.99
Seasonal Naive	262.937	213.12	172.89	73.11
ARIMA	283.996	225.77	76.12	68.37
SARIMA	205.786	153.28	12.57	34.42

Table 2. Performance Metrics for the ICD-10 Department Prediction Models

Table 2: Prediction Model Accuracy - ICD-10 Department	AIC	RMSE	MSE	MAE
Simple Naive	655.772	371.46	138.98	78.93
Seasonal Naive	359.801	171.68	97.76	46.82
ARIMA	389.844	142.91	88.43	58.76
SARIMA	147.394	134.59	85.44	42.65

Table 3. Performance Metrics for the ICD-10 Respiratory Virus Prediction Models

Table 3: Prediction Model Accuracy - ICD-10 Respiratory Viruses	AIC	RMSE	MSE	MAE
Simple Naive	878.114	395.47	203.34	154.45
Seasonal Naive	470.512	215.65	109.65	53.24
ARIMA	482.218	289.43	51.27	51.46
SARIMA	443.752	82.39	43.76	35.18

Bibliography

1. Dong, E., H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time.* Lancet Infect Dis, 2020. **20**(5): p. 533-534.

2. Syeda, H.B., et al., *Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review.* JMIR Med Inform, 2021. **9**(1): p. e23811.

3. Li, K., M. Al-Amin, and M.D. Rosko, *Early Financial Impact of the COVID-19 Pandemic on U.S. Hospitals.* J Healthc Manag, 2023. **68**(4): p. 268-283.

4. United States. Congress. Senate. Committee on Banking Housing and Urban Affairs, *The quarterly CARES Act report to Congress : hearing before the Committee on Banking, Housing, and Urban Affairs, United States Senate, One Hundred Sixteenth Congress, second session, on examining testimony from the Secretary of the Treasury and the Chairman of the Federal Reserve, as required under Title IV of the CARES Act : December 1, 2020*. S hrg. 2021, Washington: U.S. Government Publishing Office. iii, 80 pages.

5. Lalani, K., et al., *The Impact of COVID-19 on the Financial Performance of Largest Teaching Hospitals.* Healthcare (Basel), 2023. **11**(14).

6. Taubenberger, J.K. and D.M. Morens, *1918 Influenza: the mother of all pandemics.* Emerg Infect Dis, 2006. **12**(1): p. 15-22.

7. Sharma, R.R., et al., *EVDHM-ARIMA-Based Time Series Forecasting Model and Its Application for COVID-19 Cases.* IEEE Trans Instrum Meas, 2021. **70**: p. 6502210.

8. Duangchaemkarn, K., et al., *SARIMA Model Forecasting Performance of the COVID-19 Daily Statistics in Thailand during the Omicron Variant Epidemic.* Healthcare (Basel), 2022. **10**(7).

9. Mahony, J., et al., *Development of a respiratory virus panel test for detection of twenty human respiratory viruses by use of multiplex PCR and a fluid microbead-based assay.* J Clin Microbiol, 2007. **45**(9): p. 2965-70.

10. Prevention, C.f.D.C.a. *Background for CDC's Updated Respiratory Virus Guidance*. 2024; Available from: https://www.cdc.gov/respiratory-viruses/background/index.html.

11. Prevention, C.f.D.C.a. *Infection Control Guidance*. 2023; Available from: https://www.cdc.gov/coronavirus/2019-ncov/hcp/infection-control-recommendations.html.