Mapping the Seasons of Sickness:

Geographic and Seasonal Trends in U.S. Disease Incidence

Bryan Jacobs, Partha Vemuri, Yash Sharma, and Antonio Escalante Jr.

University of Arizona

Introduction

The primary objective of this report is to analyze the historical trends and patterns of disease incidence in the United States using data from the "Project Tycho Level 1 Data" dataset acquired from healthdata.gov. This dataset provides weekly reports of epidemiological data across multiple diseases dating back to the early twentieth century, enabling an exploration of time-series trends, geographic variations, and seasonal impacts. The research focuses on predicting high-risk locations and times of year in the United States for the measles virus. 283 measles cases were reported in the United States alone in 2024. Measles was the focus disease of the research because of ongoing measles outbreaks worldwide; especially in regions with low vaccination rates (Centers for Disease Control and Prevention [CDC], n.d.). Insights derived from this analysis aim to guide public health interventions and resource allocation in the event of a major outbreak in the United States.

Methods

Data

The dataset contains 759,467 observations, each representing weekly disease incidence data at the state level. Key variables include “epi_week”, representing the epidemiological week in a YYYYWW format, and “state”, which abbreviates the location. The “loc” variable provides the full name of the location, and “loc_type” specifies the type of location. The dataset includes “disease”, which indicates the disease name, “cases”, which records the number of reported cases, and “incidence_per_100000”, representing the incidence rate of each disease per 100,000 population.

Data preprocessing included converting “epi_week” to a datetime format for temporal analyses. Additional transformations were applied to align correctly with the logistic regression model, including the extraction of week and year components from “epi_week”. A “season” variable was derived to categorize weeks into winter, spring, summer, or fall. A “risk” variable was created based on the 75th percentile of “incidence_per_100000”, classifying the risk as high or low.

Exploratory data analysis examined time-series trends and geographic variations. Time-series plots were generated to visualize incidence rates over time for major diseases, highlighting historical outbreaks and vaccination impacts. Geographic disparities were mapped, showcasing average incidence rates by state using choropleth and bubble maps. Seasonal patterns were analyzed by aggregating incidence rates across seasons to identify periods of heightened disease activity.

Data Analysis

A logistic regression model was developed to predict high-risk seasons and locations for measles. After attempts with linear regression and k-means clustering, logistic regression was determined the best model choice. Linear regression was unusable due to the non-linear relationship between the independent and dependent variables. K-means clustering produced useful results with sufficient post-hoc test scores, but we found logistic regression to exhibit the perfect balance of performance and presentability to any audience.

The predictors in the logistic regression included state, encoded as dummy variables with Arizona as the baseline, and season, encoded as dummy variables with fall as the baseline. The response variable was the binary classification of risk level. A stepwise regression was not necessary as the desired predictor variables were predetermined for the purpose of the research. The dataset was split into training and testing sets, with an 80-20 split. Model performance was evaluated using accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC).

Results

The logistic regression model successfully classified high-risk seasons and locations for measles, achieving a respectable performance with an AUC (Area Under the Curve) value of 0.73. This indicates that the model was notably better than random guessing (AUC = 0.5). The model was also evaluated using accuracy, precision, recall, and F1-score. The model's accuracy of 0.67 indicates that the model correctly predicted the risk of cases 67% of the time. The model's precision of 0.64, recall of 0.59, and F1-score of 0.62 indicate a moderate balance of identifying positives and avoiding false positives. The results highlight clear seasonal and geographic patterns in the risk of contracting measles, providing insights into factors influencing disease prevalence.

Seasonal Risk Patterns

Spring emerged as the season with the highest risk of measles. The odds of being at high risk for measles in the spring were approximately 6.5 times greater than in the fall, which served as the baseline season as well as the season of lowest risk. This significant seasonal effect aligns with historical epidemiological patterns, where measles outbreaks have been more prevalent during specific times of the year.

Geographic Risk Patterns

Geographic disparities in measles risk were evident in the analysis. Wisconsin was identified as the state with the highest risk, while Louisiana exhibited the lowest risk. Comparatively, individuals moving from Arizona (the baseline state) to Wisconsin faced odds of being at high risk that were approximately 2.8 times greater. Conversely, moving from Arizona to Louisiana reduced the odds of being at high risk to approximately 0.3 times. These findings underscore the influence of location on disease prevalence and may reflect variations in vaccination coverage, public health interventions, and population density.

Discussion

The identification of seasonal and geographic patterns in measles risk has significant implications for public health strategies. The increased risk in spring highlights the need for targeted awareness campaigns and vaccination drives ahead of the season to mitigate potential outbreaks. Geographic insights further support resource allocation, enabling states with higher risk, such as Wisconsin, to prioritize immunization and preventive measures.

While the model performed well in distinguishing high-risk locations and seasons, there is room for further improvement. Validation with additional datasets could strengthen the reliability of the results and ensure the model generalizes well to other diseases and contexts. Moreover, expanding the analysis to include other diseases would provide a more comprehensive understanding of disease prevalence across the United States, enhancing the utility of the approach for public health planning.

Future research should focus on refining the classification thresholds for risk levels, exploring interactions between predictors, and investigating the impact of demographic factors such as age and population density. Such extensions would contribute to a deeper understanding of the factors driving disease outbreaks and help develop more robust models to guide public health decision-making.

References

"Measles Cases and Outbreaks." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/measles/data-research/index.html. Accessed 6 Dec. 2024.

"Project Tycho ® Level 1 Data - G89T-X93H - Archive Repository." *HealthData.Gov*, 18 Nov. 2024, healthdata.gov/dataset/Project-Tycho-Level-1-Data-g89t-x93h-Archive-Repos/i7ve-chx4/about_data.