Project Title:
Risk Profiling Unvaccinated Children for RSV in Orange County

Student Names
Ifigeneia Stathaki – 47924477, istathak@uci.edu
Arun Premanand – 59506594, apremana@uci.edu
Aryaman Tyagi – 34325612, aryamat@uci.edu
Ekaterina Kladova – 53568374, ekladova@uci.edu
Sandra Ossman – 19145203, sossman@uci.edu
Vipin Chinnaswamy – 39474565, vchinnas@uci.edu

GitHub Repository And/Or Links to Shareable Resources:
Due to CHOC's protocol, analysis was not possible to make public on GitHub. Google Drive files are provided instead.

Dashboard Zip File (anyone with UCI email can view):
https://drive.google.com/file/d/1NnBTnzH3Il4HO6vWvmU-kuPpDZdbTN94/view?usp=sharing

Code Files (anyone with UCI email can view):
https://drive.google.com/drive/folders/1W4sdlOenYMaq8N9jV2NzNnFXobqGAIA6?usp=sharing

Introduction and Problem Statement

Respiratory Syncytial Virus (RSV) is a flu-like respiratory virus that represents a major public health challenge, accounting for a substantial proportion of pediatric hospitalizations in children under 24 months old. Children's Hospital of Orange County (CHOC) serves the community in Orange County, and wanted to know and predict the risk of RSV in children. Because of the focus on Orange County, CHOC was interested in zipcode-level insights and instructed us to ignore patients outside of the area. They provided us with patient health histories of all patients who had an RSV diagnosis, but were unable to provide a control group (workarounds discussed in the Technical Approach Section).

Our deliverables included a dashboard where non-technical stakeholders could easily look up trends across Orange County and an Ordinal Regression model which predicted the severity of a patient's RSV based on a small number of features.

Related Work

Our sponsors at CHOC provided us with some published work around RSV and epidemiology modeling. *Game-Theoretic Frameworks for Epidemic Spreading and Human Decision-Making: A Review* (Huang and Zhu) shows that disease transmission dynamics are influenced by individual behavioral incentives and strategic health responses to perceived risk. *Cohort profile: A population-based record linkage platform to address critical epidemiological evidence gaps in*

respiratory syncytial virus and other respiratory infections (Sarna et al.) demonstrates integration of multi-source health datasets to characterize RSV burden, outcomes, and risk factors at scale. *Negative network effects and public policy in vaccine markets* (Amir, Liu, and Tian) explains how behavioral and informational externalities can suppress vaccine uptake and shift population-level susceptibility.

Almogy et al. introduced the concept of *Local Transmission Zones (LTZs)* to show that influenza and RSV circulate within semi-independent micro-communities rather than a single homogeneous population. By applying k-means clustering to geocoded patient home locations, they identified spatially coherent LTZs where temporal patterns of each virus showed minimal overlap and stronger single-pathogen dominance.

Surie et al. (2025) examined demographic, clinical, and community-level factors influencing RSV vaccine uptake across 20 U.S. states using FluSurv-NET data. They found higher vaccination among individuals with chronic or immunocompromising conditions and lower uptake among those with Medicaid, no insurance, or living in socially vulnerable areas. Together, these works motivate RSV risk profiling and clustering frameworks that draw on comprehensive linked patient encounter data.

Building on this idea of uncovering hidden structure, our work shifts the focus from spatial transmission zones to patient-level phenotypes. Using K-Prototypes clustering, we segment RSV-positive patients based on demographics, comorbidities, and encounter characteristics to identify distinct clinical risk groups.

Data Sets and Exploratory Data Analysis

The RSV dataset includes 10,731 rows with 10,070 unique patients and encounters, covering July 2012 to August 2025. The dataset has very few missing values overall, with small gaps appearing mainly in ICU class, reason for visit, and length of stay. There is no structural missingness, and the small number of nulls likely reflects incomplete documentation for older encounters or outpatient cases with shorter stays.

RSV records rise sharply after 2020, reaching the highest count in 2022. The numbers then decline in 2023–2025 (Appendix Fig 3). The dataset is mostly White and Other race categories. Smaller groups like Asian, Black, and others appear in much lower counts (Appendix Fig 5). Infants under 1 year make up the largest group, followed by younger children. Older age groups show steadily decreasing counts (Appendix Fig 4).

The chart shows that RSV cases peak in the winter months and drop to very low levels during the summer. This clear seasonal pattern shows a sharp rise in late fall and early winter, with the lowest levels in summer (Appendix Fig 6).

The dataset also included ICD codes, which are used for labeling diseases, symptoms, and medical conditions in healthcare records. Another column that proved useful was the "highest respiratory support," which was calculated by our industry partner to specify which of the treatments that the patient received was the most intensive.

Technical Approach / Methods

Due to the data's lack of a control group, and the lack of publicly available data on Orange County's population by zipcode, there was no way to calculate the risk of a child getting RSV, as per CHOC's original request. Instead, we redefined "risk" of RSV as "severity," since some children handle the infection better than others. However, "severity" does not have a formal definition, though some clinicians have made attempts at measuring it (Caserta, Zakariya). However, the data did not include many of the variables that these papers' measures rely on, and so we were unable to implement any of them exactly.

Therefore, a new measure of severity needed to be made. With this, we had three approaches:

- Approach 1: using "highest respiratory support" column as a tiered ranking,
- Approach 2: clustering patients using K-Prototypes,
- Approach 3: using PCA and then clustering

After testing all three methods, Approach 1 was chosen for its simplicity and explainability, since CHOC valued being able to explain the decisions their clinicians were making. An ordinal regression model was built using these tiers as the targets and achieved an accuracy of 58%. Further details are discussed in the Ordinal Regression section. The K-Prototypes model was added to the dashboard to help visualize the geographical spread of patients in each cluster.

Cluster Data Preprocessing

Since CHOC was specifically interested in infants and toddlers, we first subsetted our dataset to only patients under the age of 2 years old. Patients may visit the hospital multiple times, so we made sure to only focus on their first visit to capture only initial findings and diagnoses. We also excluded columns that related to the final RSV diagnosis, since we did not want target values leaking into the model's decisions.

We had to fill in some missing values for columns such as length of stay days, length of stay hours, and reason for visit. For numerical columns we used median or mean, and for categorical columns we filled with 0 or Unknown. For our numerical features, we used a StandardScaler from sklearn to make sure that distances are not blown up when performing clustering.

Approach 1: "highest respiratory support" tiers

As discussed previously, the data was preprocessed by the medical team to have a column of the highest form of treatment that the patient received. By observing which treatment levels that the patient received (e.g. has_intubation True/False, has_CPAP True/False) got overridden in the

"highest respiratory support" column, we were able to determine the relative rankings of the treatments, and thus use the seven categories as an indicator of severity (see figure to the left). However, using all seven tiers separately lead to a low accuracy from the clustering models. Therefore, we combined the tiers based on how easy the treatment is to administer: low severity (room air, nasal cannula, or low-flow oxygen), moderate severity (high-flow nasal cannula, CPAP, or BiPAP), and high severity (intubation or endotracheal tube ventilation).

Approach 2: Clustering with K-Prototypes

K-Prototypes Algorithm

K-Prototypes is a clustering algorithm designed specifically for datasets that contain a mixture of numerical and categorical features. Traditional clustering methods like K-Means work only with numerical variables because they rely on Euclidean distance, while K-Modes k ,works only with categorical variables because it uses simple matching of category levels. K-Prototypes combines both the models into a single unified architecture so it can handle mixed type data in a meaningful and interpretable way. The core of the K-Prototypes architecture is its hybrid distance function. It merges numerical distance and categorical dissimilarity into one combined metric (Appendix Fig 11-12 for in-depth explanation).

Assignment and Update Steps

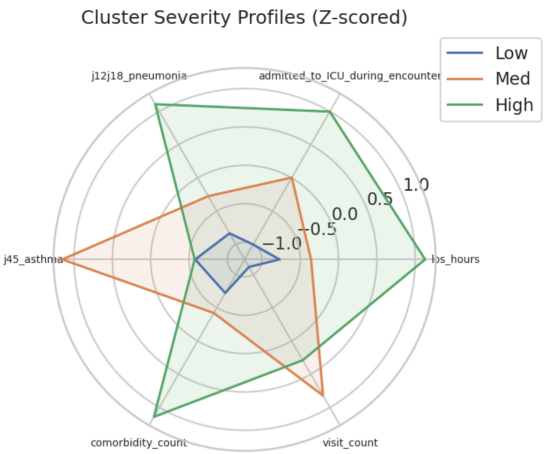
In each iteration, every data point is assigned to the cluster whose centroid has the smallest hybrid distance that combines numeric differences and categorical mismatches. After assignments are made, the centroids are updated. Numerical features are updated using the mean, and categorical features are updated using the mode. These updated values form the new centroids, and the process repeats until the clusters stop changing or the maximum number of iterations is reached.

Cluster Features

Understanding respiratory illness in hospitalized children requires more than just clinical data, it also needs an understanding of the real-world challenges clinicians face when symptoms overlap, seasons collide, and testing decisions must be made quickly. For this we studied prior work such as Clinical Testing for COVID-19, Influenza, and RSV in Hospitalized Youths (Toepfer) which highlighted to us that RSV and influenza often surge at the same time, and their symptoms can look nearly identical in infants. As a result, many children are tested for multiple pathogens at once, making it difficult to cleanly distinguish RSV only patients from those experiencing co-infections. Alongside this, we also studied pediatric respiratory illness disparities and the work done in Systematic Literature Review of Risk Factors for Poor Outcomes Among Children With Respiratory Syncytial Virus Infection (de Vries) showed us that factors like age,

race/ethnicity, and insurance type can meaningfully shape severity, length of stay, and the probability of ICU admission.

We wanted to separate the patients into Low, Med, and High Risk/Severity clusters so we set K=3 for the K-Prototypes model. We looked at clinically interpretable statistics such as mean age at encounter, average length of stay, ICU indicators, maximum visit counts, and 30-day readmission risk to see what defined the clusters. Demographic variables such as sex, race, ethnicity, and insurance status were used because prior research (de Vries) consistently showed us these attributes are strongly tied to access to care and clinical outcomes. For infection status, we used PCR confirmed RSV results, which allowed us to cleanly separate RSV only patients from those with broader viral involvement, reflecting the approach recommended in recent clinical surveillance papers such as Risk Factors for Severe Disease Among Children Hospitalized With Respiratory Syncytial Virus (Bont).



The resulting clusters seemed to separate very well across the visit metrics and comorbidities as seen above. Length of stay hours seemed to be the most variable characteristic across the clusters, as well as ICU admittance, which is to be expected. The visit counts were very close to 1 across the clusters, which means that RSV was able to be diagnosed on the first visit (Appendix Fig 7). Interestingly, nearly all the asthma patients were separated out into the Med Risk cluster, as seen in the web plot to the left.

Approach 3: PCA and then clustering

To address the issue of clustering methods not performing well with high-dimensional data, we also ran a PCA to determine the most influential columns by summing the absolute loadings of the components for each column. This gave us a ranked list of features. Then, we took the top 15 of those features (Appendix Fig 1) and ran a KMeans algorithm on them (not K-Prototypes since all columns were numerical) to group the data into 3 clusters (again aiming for low, medium, and high). From there, we looked at the average values of each column to identify the ones with the most disparity between the clusters (Appendix Fig 2). These columns could be used to identify the clusters and find meaning for why the algorithm ended up putting them together. They seemed to follow a similar pattern, where one group was relatively healthy and the others were less so.

This approach is more hands-off than manually selecting which columns to cluster by, and which columns to define clusters and thus avoids any human bias that might influence in the analysis. However, it was a lot more complicated and thus harder to explain, and so while it helped confirm our findings about the three clusters, it was not used as the final way of clustering.

Ordinal Regression

While clinicians can assess disease severity through examination, real-time automated risk stratification tools remain scarce. This portion of the project addresses this gap by developing an ordinal severity classification model to predict respiratory support that will be needed by the patients across three levels: Low, Moderate, and High. By treating severity as an ordered outcome rather than discrete categories, the ordinal regression appropriately penalizes distant misclassifications more heavily than adjacent ones. For example, low misclassified as high should have higher penalty than moderate classified as high. This effect is highly useful in medical settings. The resulting model aims to identify the most predictive clinical variables and provide a foundation for enhancing a provider's ability to mitigate patient risk and properly allocate resources.

The proportional odds model, first introduced by McCullagh (1980), has become a standard framework for such problems, as it respects the inherent ordering of outcomes while avoiding the limitations of treating severity as either a continuous metric or a set of independent categories. Specific studies have looked at ordinal regressions for hospital admission and risk admission; however, these studies have focused on older adults and not young children (Suarez-Betancourt).

Data Cleaning

The primary outcome variable was severity, and was operationalized as the maximum level of respiratory support required during the encounter. Three ordered categories were defined: low, moderate, and high severity. This ordinal structure reflects the clinical progression of respiratory distress and aligns with standard escalation protocols in pediatric emergency and inpatient care.

Patient identifiers and variables associated with treatment decisions (e.g., ICU admission, procedures performed) were removed to prevent data leakage, as these outcomes occur after or concurrent with the severity determination and should not inform prediction. To address multicollinearity among diagnostic and comorbidity features, variance inflation factor (VIF) analysis was applied with a threshold of 10; features exceeding this threshold were iteratively removed or collapsed, where applicable, until all remaining predictors exhibited acceptable collinearity. Following feature selection, continuous variables were standardized using z-score normalization, with scaling parameters fit on the training set and applied to the test set to prevent information leakage. Finally, to mitigate class imbalance, particularly the underrepresentation of High severity cases, random oversampling was applied to the Moderate and High severity classes in the training set, ensuring the model received sufficient signal from minority classes during optimization.

Architecture

We employed an ordinal logistic regression model using the LogisticAT implementation from the mord Python library, which applies the proportional odds assumption with L2 regularization to control overfitting. The regularization strength hyperparameter, α, was tuned via grid search over five values spanning 0.01 to 10.0, with model performance evaluated on a held-out validation

subset. An α value of 1.0 was selected based on optimal validation accuracy and macro-averaged F1 score. Model performance was assessed using an 80/20 stratified train-test split, preserving the severity class distribution across subsets. Primary evaluation metrics included overall accuracy and macro-averaged F1 score, the latter chosen to account for class imbalance by weighting each severity level equally regardless of prevalence. Macro-F1 is particularly appropriate in clinical contexts where sensitivity to minority classes (high severity) is critical. To quantify the impact of the feature, we conducted a McNemar test comparing the full and reduced models on paired predictions from the test set. This non-parametric test evaluates whether the two models differ significantly in their classification errors, providing statistical evidence for or against the value of the excluded feature.

Results

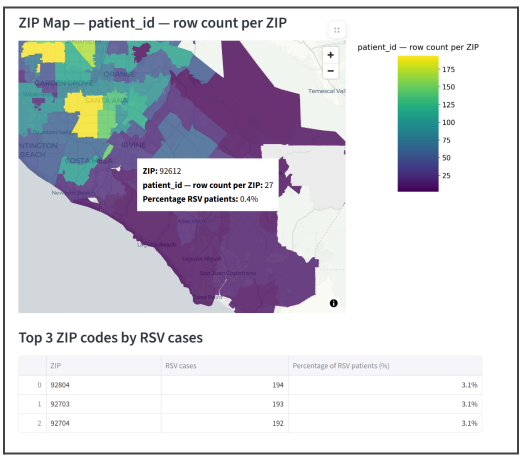
The full ordinal logistic regression model achieved an overall accuracy of 0.581 and a macro-averaged F1 score of 0.501 on the held-out test set . The nested model, which excluded the j_other_resp_dx feature, demonstrated reduced performance with an accuracy of 0.563 and macro-F1 of 0.475. McNemar's test comparing the two models on paired predictions yielded a p-value of 0.1125, indicating a trend toward significance but failing to reach the conventional ($p < 0.05$) threshold. This may suggest that while j_other_resp_dx contributes predictive signal, its impact is moderate and may reflect redundancy with other respiratory diagnostic features.

Final features chosen for the predictor model included clinical utilization metrics (PCR positivity, prior hospital visits, and visit frequency), respiratory-related ICD-10 diagnosis codes (J-code groupings), and comorbidity indicators spanning cardiac, hematologic, neurologic, metabolic, and other diagnoses. These features were selected based on clinical relevance and sponsor conversations suggesting that baseline health status and acute presentation characteristics are the key drivers of RSV severity.

Software and Codebase

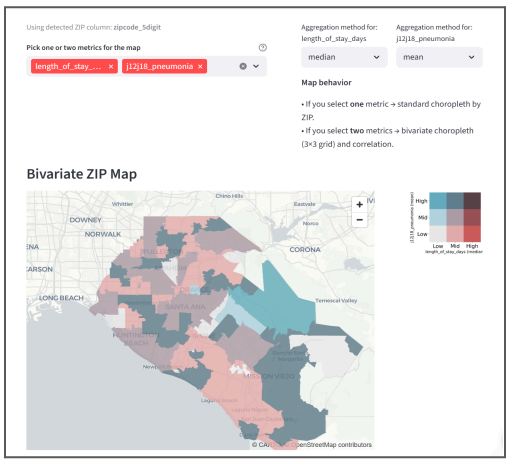
The RSV web app Dashboard is an interactive geospatial analytics tool designed to support clinicians and public-health professionals in exploring RSV-related patterns across Orange County at the ZIP-code level. By integrating patient encounter data and demographic information with geographic boundaries, the system enables rapid and intuitive visualization of community-level trends. This section summarizes the components of the dashboard.

The primary component of the web app is the ZIP Code Risk Profiler, which presents a choropleth map (pictured to the right) that allows users to visualize any selected clinical or demographic metric, such as pneumonia incidence, ICU admissions, length of stay, or social vulnerability indicators. Users may choose from several aggregation methods including mean, median, standard deviation, and count, and the system dynamically computes



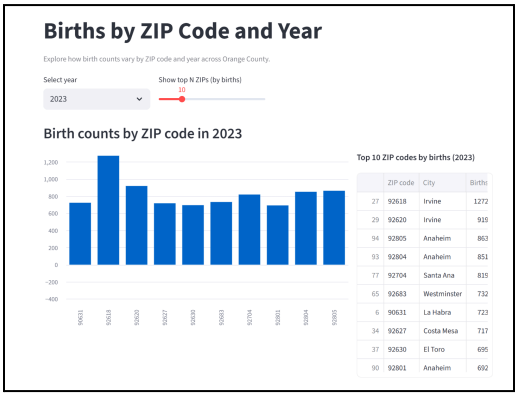
ZIP-level summaries. Hover-over tooltips provide additional context by displaying both the aggregated value and the number of patient encounters associated with each ZIP code. This functionality enables clinicians to quickly identify geographic hot spots, detect uneven distributions of illness severity, and understand where patient volume is concentrated, supporting more accurate situational awareness and interpretation.

The dashboard also includes a bivariate metric comparison page, shown on the right, that allows users to visualize two metrics simultaneously through side-by-side choropleths, an integrated bivariate map, and an automatically generated correlation matrix. This feature helps clinicians examine how clinical outcomes relate to contextual or social factors, such as patient age and severity clusters. By enabling exploration of co-occurring risk factors, this page supports more targeted intervention planning, informed resource allocation, and a deeper understanding of the underlying drivers of RSV burden.



Across both mapping pages, the system incorporates an aggregation engine that computes descriptive statistics for selected ZIP codes, with patient count serving as a particularly important indicator of sample size and population density. Including count helps clinicians distinguish between trends arising from large, clinically meaningful populations and those driven by small or unstable samples, thereby improving the reliability of their interpretations.

The dashboard also provides a dedicated Birth Rate Trends page that visualizes annual birth rates at the ZIP-code level. Because infants represent the population most vulnerable to severe RSV outcomes, understanding where newborn populations are concentrated or increasing is essential for anticipating future disease burden. This feature aids clinicians and health systems in planning pediatric capacity, guiding preventive efforts, and identifying areas where increased outreach may be beneficial.



Lastly, the final page in the application incorporates the ordinal severity risk model developed and finalized in our technical approach. On the “RSV Severity Predictor” page, clinicians enter patient-level characteristics (e.g., age at encounter, ICD-10–based comorbidities, prior visits, and selected social risk indicators), which are assembled into a feature vector, scaled, and passed to an ordinal logistic regression model. The tool returns a predicted severity category (Low, Medium, or High) together with the estimated class probabilities, and it also displays summary performance metrics (accuracy, macro F1, and class-specific recall) computed on a held-out test set. This interface provides a transparent, point-of-care decision-support tool that standardizes

risk estimation for infant RSV encounters while explicitly communicating model limitations, and is intended to complement rather than replace clinical judgment.

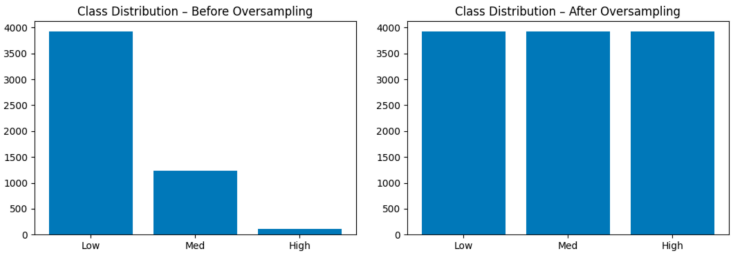
Together, these components create a unified platform for geospatial and patient-level analysis of RSV-related data. By translating complex datasets into accessible visualizations, the dashboard equips clinicians with the tools needed to identify high-risk neighborhoods, monitor community-level disease burden, and make timely, evidence-based decisions. This enhances the ability to deliver equitable and proactive care across Orange County.

Evaluation/Validation

Models were trained using an 80/20 stratified split, with all preprocessing (scaling and oversampling) restricted to the training set to prevent leakage. A small grid search on the training/validation split selected $\alpha = 1$ for the ordinal logistic regression. All results reported reflect true held-out test performance. Oversampling medium and high cases substantially improved recall for underrepresented classes (Medium: 0.27→0.65; High: 0.34→0.76). The final model achieved 0.581 accuracy and 0.501 macro-F1 on the test set, with most errors occurring between adjacent severity levels. Class-wise performance is strongest for low severity and moderate but clinically meaningful for medium and high.

K-Prototypes clustering using Gower distance produced a silhouette score of 0.229, indicating modest but present structure in patient profiles. These clusters were primarily used to explore heterogeneity in early presentations rather than for predictive modeling.

Moreover, looking at the main driver for the ordinal prediction, we decided to test nested models to evaluate the need for that feature. Ablation reduced performance (accuracy: 0.581→0.563, macro-F1: 0.501→0.475), suggesting moderate predictive value. McNemar's test ($p = 0.1125$) showed no statistically significant difference, but indicated a trend favoring the full model. The feature appears useful overall.



Recall Before	Recall After
Medium : 0.27	Medium : 0.65
High : 0.34	High: 0.76

Confusion matrices and per class metrics showed most errors involve borderline medium cases, consistent with clinical ambiguity. Oversampling mitigated the model's initial bias toward predicting low severity and improved balance across classes. Coefficient plots and class distributions were used to verify that the model remains interpretable and clinically coherent (figure below) (Appendix Fig 10).

Per our sponsors and given the current dataset structure, the most rigorous validation available is randomized N-fold cross-validation at the patient level, which fairly assesses generalization for

our “flat” single-encounter model. However, the ordinal logistic regression model using mord with L2 regularization and repeated resampling within each fold was too computationally heavy for this project.

Team Member Participation

Arun, Aryaman, and Vipin primarily worked on the K-Prototypes clustering model. Sandra spearheaded the Streamlit dashboard with help from Arun and Ifi, who added the parts related to their models. Sandra, Ifi, and Katya were the ones to send emails and take meeting notes. Regression was done primarily by Ifi, with help from Katya. Everyone consistently came to meetings and helped with writing the report/making presentations.

Clustering	33% - Arun	33% Aryaman	33% - Vipin
ZipCode Dashboard	10% - Arun	75% - Sandra	15% Ifi
Notes and Communications	50% - Sandra	25% - Ifi	25% - Katya
Regression	75% - Ifi	25% - Katya	

Discussion and Conclusion

This project revealed several important insights about working with clinical RSV data. HIPAA constraints made it difficult to access detailed patient-level information, and many useful variables were unavailable or highly limited. ZIP-code data, including birth rates, were sparse, which pushed us to rely on broader aggregates rather than more granular analyses. As a result, much of our work focused on making the best use of limited features while still producing outputs that were clinically meaningful.

Our approach had clear limitations. Without a control group of non-RSV patients, our predictive modeling cannot isolate RSV-specific risk factors and has limited ability to generalize. The small feature set also restricts model performance and reduces interpretability. A major challenge was telling a coherent story with the data available; designing visualizations and selecting metrics required careful judgment to avoid overstating patterns that may be driven by small sample sizes or missing information.

If more time were available, creating a proper control group and expanding the feature set would be the most impactful next steps. Additional work could include building a more robust patient-level severity predictor, incorporating temporal trends to capture RSV seasonality, or integrating external datasets such as environmental or socioeconomic indicators. These extensions would strengthen the dashboard’s predictive capabilities and increase its value for clinicians planning for RSV seasons.

Works Cited

Almogy, G., Stone, L., Bernevig, B. A., Dorozko, M., Wolf, D. G., Moses, A. E., & Nir-Paz, R. (2017). Analysis of influenza and RSV dynamics in the community using a 'Local Transmission Zone' approach. *Scientific Reports, 7*, Article 42010. https://doi.org/10.1038/srep42010

Amir, R., Liu, Z., & Tian, J. (2023). Negative network effects and public policy in vaccine markets. *Journal of Economic Behavior & Organization, 208*, 140–155.

Bont, N. J., et al. (2023). Risk factors for severe disease among children hospitalized with respiratory syncytial virus.

Caserta, Mary T et al. "Development of a Global Respiratory Severity Score for Respiratory Syncytial Virus Infection in Infants." *The Journal of infectious diseases* vol. 215,5 (2017): 750-756. doi:10.1093/infdis/jiw624

de Vries, X. J., et al. (2023). Systematic literature review of risk factors for poor outcomes among adults with respiratory syncytial virus infection in high-income countries. *Clinical Infectious Diseases*.

Huang, Y., & Zhu, Q. (2020). *Game-theoretic frameworks for epidemic spreading and human decision-making: A review. Dynamic Games and Applications, 10*(3), 844–883.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1980.tb01109.x

Paramo, M. V., Watts, A. W., Bone, J. N., et al. (2022). RSV hospital admissions during the first 2 seasons among children with chronic medical conditions.

Sarna, M., Taye, B., Le, H., Giannini, F., Glass, K., Blyth, C. C., Richmond, P., Glauert, R., Levy, A., & Moore, H. C. (2022). Cohort profile: A population-based record linkage platform to address critical epidemiological evidence gaps in respiratory syncytial virus and other respiratory infections. *International Journal of Epidemiology, 51*(3), e55–e64.

Suarez-Betancourt, Lucia and Najera, Alberto and Gómez-Juárez Sango, Ana Gómez-Juárez Sango and Cantero Escribano, José Miguel and Robles Fonseca, Lorena and Simarro Cordoba, Encarnacion and García Guerrero, Jesús and Gonzalez-Rubio, Jesus, Risk Factors for Hospital Admission, Clinical Severity, and 90-Day Mortality in Adults with Respiratory Syncytial Virus (RSV) Infection. http://dx.doi.org/10.2139/ssrn.5261622

Toepfer, A. P., Rutkowski, R., Moline, H., & Dawood, F. (2025). Clinical testing for COVID-19, influenza, and RSV in hospitalized youths, 2016–2024.

Zakariya Sheikh, Ellie Potter, You Li, et al., Validity of Clinical Severity Scores for Respiratory Syncytial Virus: A Systematic Review, *The Journal of Infectious Diseases*, Volume 229, Issue Supplement_1, 15 March 2024, Pages S8–S17, https://doi.org/10.1093/infdis/jiad436

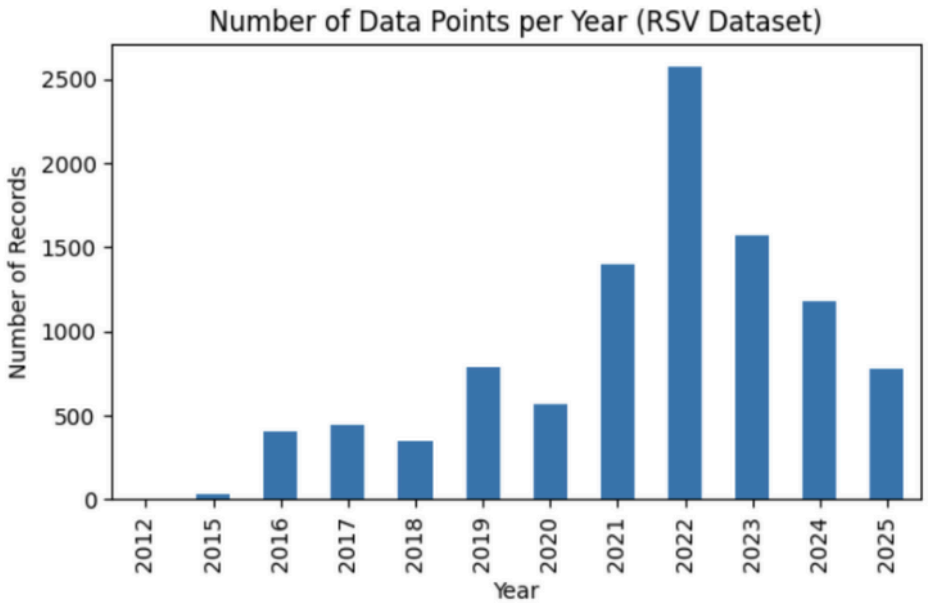
Appendix

j09j11_influenza	4.841012884102106
e70e88_metabolic_dx	4.775657966706387
e40e46_malnutrition	4.706830493132663
m_musculoskeletal_connective_dx	4.697723877149147
i60i69_cerebrovascular_dx	4.692435238892983
j45_asthma	4.6848951536053995
h00h59_ear_mastoid_dx	4.668098456058098
j12j18_pneumonia	4.639862936156969
q90q99_chromosomal_abnormalities	4.614447539850574
b90b94_sequelae_infectious_parasitic_dx	4.59544235395259
i70i89_arteries_veins_cap_lymph_dx	5.254136414869348
s00s99_t14_physical_injury_trauma	5.091398549314496
z55_z65_sdoh_dx_codes	5.027316081449827
i10i16_hypertensive_dx	5.0230640105449424
e00e36_endocrinologic_dx	4.927526411439324
d50d78_hematological_dx	4.913968580338583
d80d89_immunologic_dx	4.907741872294118
l_skin_subcutaneous_dx	4.86468802246007
b35b89_b99_fungal_and_other_infectious	4.849989764141239
h00h59_eye_adnexa_dx	4.841238114679124

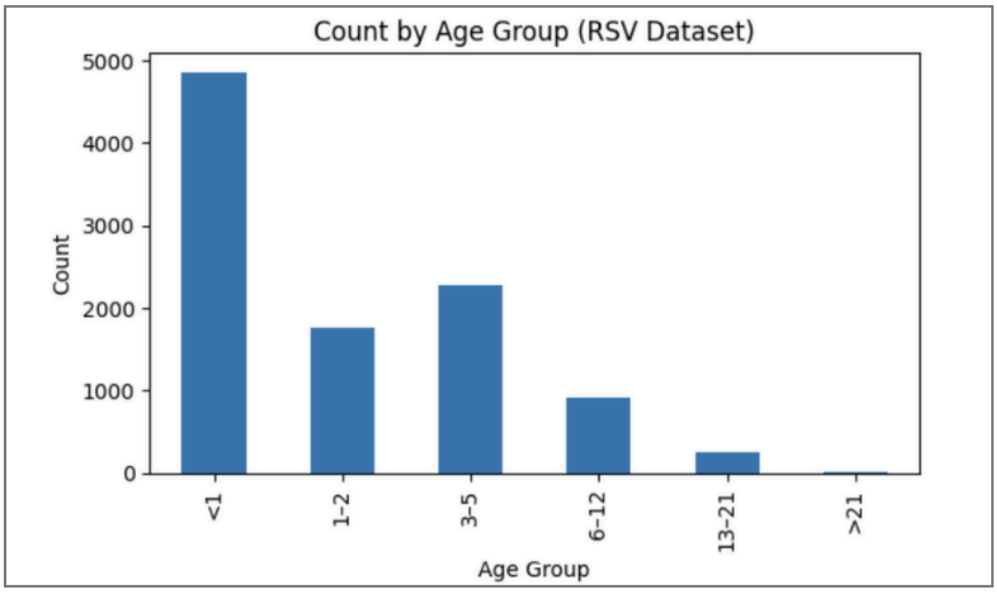
Appendix Fig 1: Top 20 most influential columns from PCA.

Cluster	0 (Low)	1 (Medium)	2 (High)
Length of stay	48 hours	113 hours	160 hours
Bacterial infections	2%	11%	14%
Other infections	1%	2%	10%
Blood conditions	3%	14%	11%
Metabolic condition	0%	100%	21%
Pneumonia	10%	25%	12%
Digestive condition	5%	11%	16%
Deep skin condition	0%	0%	100%
Birth defects	7%	14%	13%

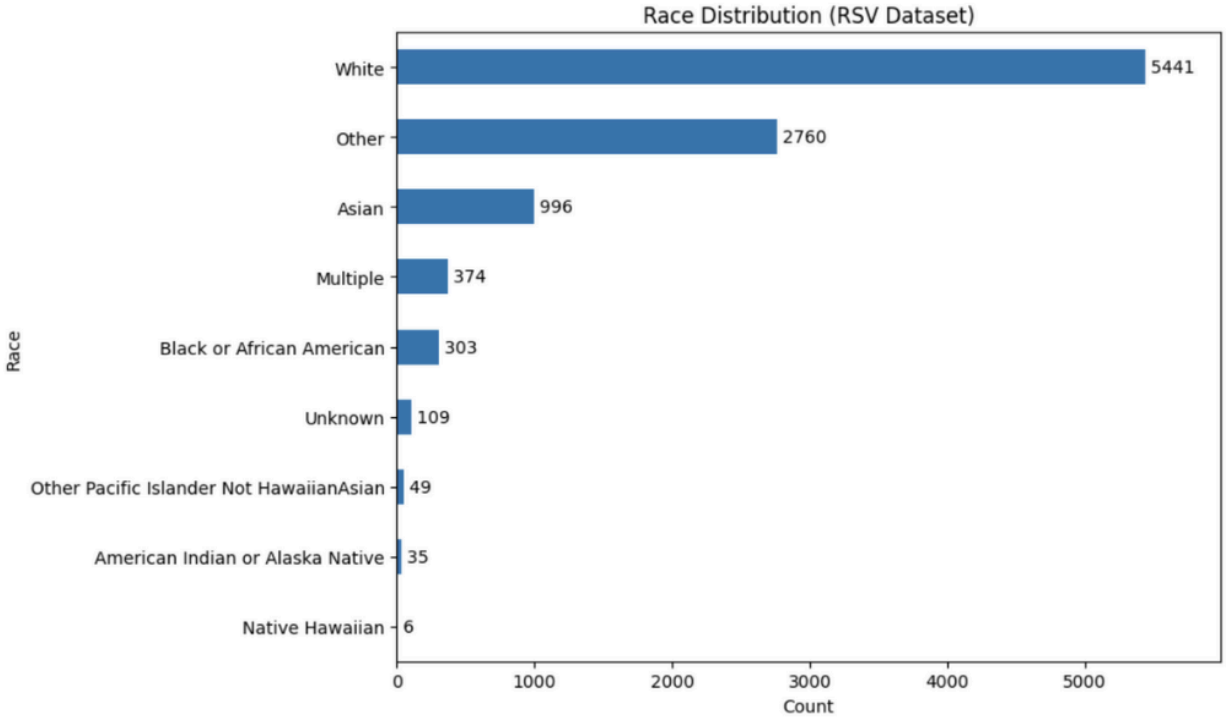
Appendix Fig 2: Columns with most disparity among clusters after clustering by columns in Appendix Fig 1.



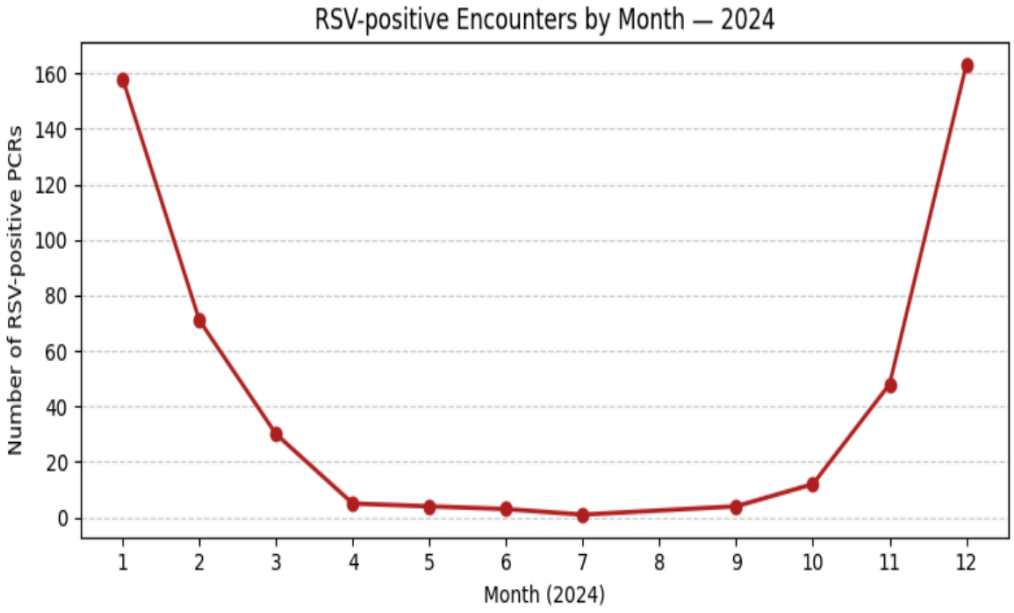
Appendix Fig 3: Distribution of data points per year.



Appendix Fig 4: Number of data points per age at encounter (hospital visit).



Appendix Fig 5: Data distribution by race.



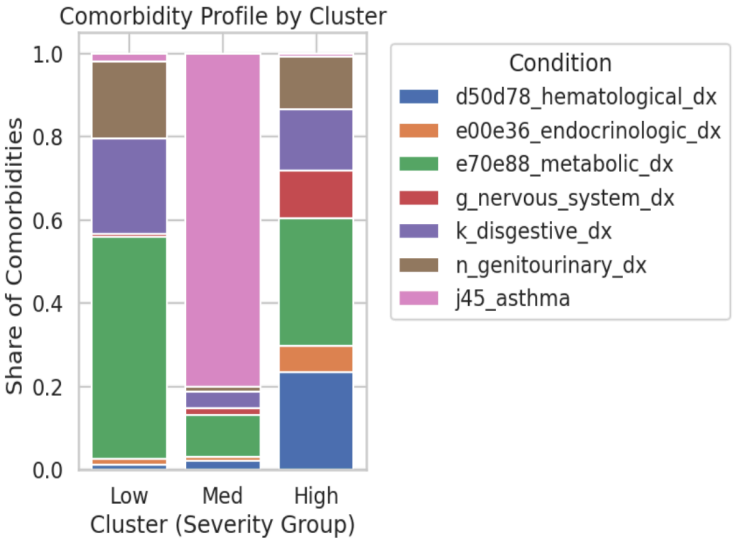
Appendix Fig 6: Datapoints (RSV related-visits) per month.

cluster	sex	race	ethnicity	insurance	age_at_encounter	los_hours	visit_count	admitted_to_ICU_during_encounter	readmit_30d_ed_or_ip
0	Male	White	Hispanic or Latino	Public	0.79	171.69	1.09	0.35	0.95
1	Male	White	Hispanic or Latino	Public	1.13	61.16	1.09	0.20	0.95
2	Male	White	Hispanic or Latino	Public	0.62	30.65	1.08	0.05	0.92

Appendix Fig 7: K-Prototypes cluster average/mode features.

cluster	0	1	2
j12j18_pneumonia	0.575181	0.166667	0.001855
a80b34_b97_viral_infections	0.484474	0.429825	0.357933
e70e88_metabolic_dx	0.282635	0.122389	0.061583
d50d78_hematological_dx	0.215685	0.026316	0.001396
a00a79_b95b96_bacterial_infections	0.184783	0.023810	0.000225
k_disgestive_dx	0.133433	0.050125	0.026463
n_genitourinary_dx	0.117650	0.012531	0.021463
g_nervous_system_dx	0.105275	0.020050	0.000655
e00e36_endocrinologic_dx	0.056126	0.010025	0.001564
j09j11_influenza	0.030083	0.017544	0.010437

Appendix Fig 8: How much of a K-Prototypes cluster has diseases related to the ICD scores.



Appendix Fig 9: Comorbidity profiles of each cluster from K-Prototypes.

Class	Precision	Recall	F1
Low (0)	0.8	0.93	0.86
Medium (1)	0.51	0.65	0.36
High (2)	0.83	0.76	0.49

Appendix Fig 10: Scores from Ordinal Regression model for classifying a data point based on given features.

$$D(i,j) = \sum_{f \in numeric} (x_{if} - c_{jf})^2 + \gamma \sum_{f \in categorical} \delta(x_{if}, c_{jf})$$

Appendix Fig 11: K-Prototypes Hybrid Distance Function for a data point x_i and a centroid c_j.

Here, the first term is the squared Euclidean distance used in K-Means. The second term which is used in K-Modes counts categorical mismatches using an indicator function that returns zero when values match and one when they differ. Gamma is a scaling factor that adjusts how strongly categorical mismatches influence the total distance. Numerical values can vary widely and may dominate the distance calculation if categorical mismatches are not amplified. Gamma solves this by scaling the categorical part so it aligns with the magnitude of numerical distances.

Appendix Fig 12: K-Prototypes Centroid Representation

In K-Prototypes, each cluster centroid is represented as a combined vector containing both numeric prototypes and categorical prototypes.

- The numeric prototype is computed as the mean of the numerical features of all points in the cluster.
- The categorical prototype is computed as the mode of each categorical feature within the cluster, meaning the most frequent category level is selected.

The centroid is therefore a hybrid object that mirrors the structure of the data itself. This dual representation ensures that both types of variables influence how clusters form.