Multi-Faceted Analysis and Prediction for the Outbreak of Pediatric Respiratory Syncytial Virus

Citation for published version:
Yang, C, Gao, J, Glass, L, Cross, A & Sun, J 2023, 'Multi-Faceted Analysis and Prediction for the Outbreak of Pediatric Respiratory Syncytial Virus', *JAMIA: A Scholarly Journal of Informatics in Health and Biomedicine (JAMIA)*, vol. 31, no. 1. https://doi.org/10.1093/jamia/ocad212

Digital Object Identifier (DOI):
10.1093/jamia/ocad212

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
JAMIA: A Scholarly Journal of Informatics in Health and Biomedicine (JAMIA)

Multi-Faceted Analysis and Prediction for the Outbreak of Pediatric Respiratory Syncytial Virus

Chaoqi Yang[1], Junyi Gao[2], Lucas Glass[3], Adam Cross[4], Jimeng Sun[1]

[1]University of Illinois Urbana-Champaign, Illinois, United States
[2]University of Edinburgh, Edinburgh, United Kingdom
[3]IQVIA, Pennsylvania, United States
[4]University of Illinois College of Medicine Peoria, Illinois, United States

Corresponding Author:
Jimeng Sun, PhD
Department of Computer Science and Carle Illinois College of Medicine
University of Illinois, Urbana-Champaign
201 North Goodwin Avenue Urbana,
Illinois, 61801, United States
Email: jimeng@illinois.edu

Abstract

Objective: Respiratory syncytial virus (RSV) is a significant cause of pediatric hospitalizations. This paper aims to utilize multi-source data and leverage the tensor methods to uncover distinct RSV geographic clusters and develop an accurate RSV prediction model for future seasons.

Materials and Methods: This study utilizes five-year RSV data from sources, including medical claims, CDC surveillance data, and Google search trends. We conduct spatio-temporal tensor analysis and prediction (TAP) for pediatric RSV in the US by designing (i) a non-negative tensor factorization (NTF) model for pediatric RSV diseases and location clustering; (ii) and a recurrent neural network tensor regression model for county-level trend prediction using the disease and location features.

Results: We identify a clustering hierarchy of pediatric diseases: Three common geographic clusters of RSV outbreaks were identified from independent sources, showing an annual RSV trend shifting across different US regions, from the South and Southeast regions to the Central and Northeast regions and then to the West and Northwest regions, while precipitation and temperature were found as correlative factors with the coefficient of determination $R^2 \approx 0.5$, respectively. Our regression model accurately predicted the 2022-2023 RSV season at the county level, achieving $R^2 \approx 0.3$ mean absolute error MAE<0.4 and a Pearson correlation greater than 0.75, which significantly outperforms the baselines with p-values <0.05.

Conclusions: Our proposed framework provides a thorough analysis of RSV disease in the US, which enables healthcare providers to better prepare for potential outbreaks, anticipate increased demand for services and supplies, and save more lives with timely interventions.

Keywords: Respiratory syncytial virus (RSV), tensor factorization, deep learning, pediatric diseases.

Introduction

Background and significance

Respiratory syncytial virus (RSV) is the most implicated virus in bronchiolitis and is a leading cause of infant hospitalizations[1], especially in developing countries[2]. Nationally, RSV bronchiolitis results in an estimated 58,000 hospitalizations and up to 500 deaths among children each year. Globally, it accounts for 1.4 million hospital admissions and 27,300 in-hospital deaths among infants under 6 months old annually[3]. The CDC reports that the RSV season in the US is typically from October to March, with peak activity in December[4]. Studies have been conducted to understand the risk factors causing different RSV trends in pediatric patients, and interventions to decrease RSV hospitalizations have been identified, including RSV immunoprophylaxis[5]. Researchers have analyzed global RSV seasonality, finding that RSV activity follows a decrease in temperature, and high humidity can increase activity in equatorial and tropical areas[6]. The timing and patterns of RSV transmission have also been studied using pre-pandemic datasets and various factors including temperature, humidity, precipitation, and maximal day-to-day temperature variation[7,8,9,10,11,12]. Recent studies have investigated the resurgence of RSV after the COVID pandemic in Singapore[13,14], West Australia[15], and New Zealand[16].

In the 2022-2023 season, RSV case counts have been exceedingly high, with a peak hospitalization rate of 4.9 per 100,000 in mid-November[17]. This rate is significantly higher than last year's mid-November hospitalization rate of 1.1 per 100,000 and the pre-COVID pandemic rate of about 0.5 per 100,000. This surge emphasizes the importance of in-depth analysis and accurate prediction of RSV trends, particularly in children, to develop effective prevention and intervention strategies.

Existing RSV prediction models and challenges

Most existing RSV prediction approaches are based on traditional statistical or machine learning methods. For example, Reis et al.[18] used Bayesian inference and built a super-ensemble model to forecast the US outbreaks of RSV. Korsten et al.[19] collected two consecutive prospective multicenter birth cohorts from June 2008 until February 2015 and used multivariate logistic regression analysis based on an existing statistical discriminative model[20] for RSV hospitalization prediction. Gebremedhin et al.[21] built a multivariable logistic regression model for predicting the hospitalization burden of RSV using a cohort (children younger than 5 years) collected between 2000 to 2012 in Western Australia.

These models showed decent performance in previous stable RSV seasons. However, the prior approaches face two major issues: (1) The annual RSV outbreak date and volumes have changed significantly during COVID[15,16], while these approaches mostly model the stable and periodic time series information before COVID (e.g., up to 2017 in Reis et al.[18]). Their accuracy can be significantly undermined by the distribution shift associated with the COVID pandemic during 2019-2022 since the RSV outbreak time and volume are significantly different from previous years; (2) Existing approaches cannot capture the underlying disparities in different regions by using only the past trend of RSV. The relative trends of other related infections (such as bronchiolitis due to parainfluenza) can also provide predictive signals. Capturing the location-based disparities and leveraging co-occurrence trends from other pediatric diseases can be challenging but beneficial. Incorporating these additional data sources into previous models might be infeasible or suboptimal.

This study proposes a hybrid approach that combines tensor factorization with deep learning techniques for *pediatric RSV time-series analysis and prediction*, named TAP-RSV. We address the above issues by utilizing non-negative tensor-based methods to extract interpretable disease and regional clustering features from multiple data resources, including medical claims data, surveillance data from CDC, and online search data from Google search trends during the past five years. With these extracted disease and regional features, we propose a recurrent neural network (RNN)-based tensor regression model that accurately estimates the county-level RSV case counts of the 2022-2023 RSV season, which outperforms the existing methods under various evaluations.

Materials and Methods

Data sources

Our study jointly analyzes data from several resources for RSV trend analysis and prediction.

Source 1 (Medical Claims Data) This data source contains county-level weekly count information for 19 pediatric diseases (RSV, bronchiolitis, adenovirus, rhinitis, viral pneumonia, coronavirus, COVID, hypoxia, hypoxemia, asthma, status asthmatics, parainfluenza, respiratory failure, respiratory distress, Upper Respiratory Infection (URI), rhinovirus, and Human Metapneumovirus (HMPV)). The data spans roughly five years in the US, which is structured as a tensor $\boldsymbol{I} \in \mathbb{R}^{N\times T\times 19}$, where $N = 2{,}334$ is the number of counties, $T = 258$ is the number of weeks in the record (from 12/30/2017 to 12/09/2022), and the last dimension of size 19 corresponds to 19 pediatric diseases. Each entry in tensor $\boldsymbol{I}$ indicates the count of a particular disease in a county over a certain week.

Source 2 (CDC RSV Surveillance data)[1] The next data source is from CDC RSV-NET project[22], which includes data from 58 counties in 12 states that participate in the Emerging Infections Program (California, Colorado, Connecticut, Georgia, Maryland, Minnesota, New Mexico, New York, Oregon, and Tennessee) or the Influenza Hospitalization Surveillance Program (Michigan and Utah). RSV-NET covers almost 9% of the U.S. population. This data source is also in weekly granular and spans roughly five years from 10/16/2018 to 01/14/2023. However, this data source only recorded the 10/01 to 04/30 season each year. Our study only uses the data portion with “children younger than 18 years of age”. We formulate the data as a matrix $\boldsymbol{C} \in \mathbb{R}^{12\times T_2}$, where $T_2 = 187$ is the number of weeks the data is collected. Each entry in $\boldsymbol{C}$ is the RSV weekly pediatric hospitalization rate per 100,000 people in one state.

Source 3 (Google RSV Trends)[2] The Google Trends data show changes in search interest for a specific region over time, expressed as a percentage of the maximum number of searches for that region within the selected time frame. The data is downloaded in weekly granularity for 51 US states over five years (from 02/04/2018 to 01/29/2023), where "Respiratory syncytial virus" is selected as the keyword search option. We formulate it as a matrix $\boldsymbol{G} \in \mathbb{R}^{51\times T_3}$, where $T_3 = 261$ indicates the number of weeks over five years. The entries in $\boldsymbol{G}$ indicate the search interest (in percentage) over each week in the state.

Source 4 (County-level static Data). The static data is at county-level, corresponding to $N =$

[1] https://www.cdc.gov/rsv/research/rsv-net/dashboard.html

[2] https://trends.google.com/trends/

2,334 counties in the claim database. It includes the distance matrix $\boldsymbol{M}_1 \in \mathbb{R}^{N\times N}$, the mobility distance matrix $\boldsymbol{M}_2 \in \mathbb{R}^{N\times N}$, demographics statistics $\boldsymbol{M}_3 \in \mathbb{R}^{N\times 14}$, overall COVID, hospitalization and vaccination statistics $\boldsymbol{M}_4 \in \mathbb{R}^{N\times 39}$. Here, $\boldsymbol{M}_1$ is the geographical distance matrix, measured by the Haversine distances between N counties. The mobility distance $\boldsymbol{M}_2$ includes the average mobility flows between N counties during 2020 and 2021. The mobility flow scores are collected from SafeGraph[23]. The demographical features $\boldsymbol{M}_3$ include populations of different age and race groups and the medical resource statistics collected from the county-level census dataset provided by[24]. The COVID, hospitalization and vaccination statistics $\boldsymbol{M}_4$ are collected from IQVIA databases[25] and JHU CSSE COVID-19 Dashboard[26]. We show the feature list in **Table ,**

Table .

Source 5 (State-level Climatology Data)[3] The climate date records are collected by the Iowa Environmental Mesonet project at Iowa State University. The webpage gives a listing of unofficial daily climate records for 361 national weather service (NWS) cooperative observer program (COOP) stations over 51 states in the US. We crawl the daily temperature (high and low in a day, in Fahrenheit) and the precipitation data (in Inch) from 01/02/2018 to 01/12/2023 by API[4].

RSV related disease clustering

RSV progression trends may be inferred using similar disease progressions as a reference. In this study, we utilized non-negative tensor factorization (NTF) to identify co-occurring trends among 19 pediatric diseases. Five-year trends of all diseases are shown in Supplementa (**Figure 1 (supp)**). We study the similarity among these diseases by using a subset of the medical claims data and applying sparse non-negative tensor factorization (NTF) [27,28] method. NTF is commonly used for extracting the low-rank structure of real-world count-based data, which decomposes a large high-dimensional tensor into low-rank factor matrices of each information aspect (Illustration can be found in Figure 1(a)). Due to space limitation, the dense mathematical formulation of NTF (including the detailed notations, objective functions, and optimization procedures) are presented in Supplementary Analysis 1.

We apply NTF on the claims data up to 08/20/2022 as shown in **Figure 1**(a). Note that clustering analysis avoids using any test information in the prediction experiment as the later prediction window begins on 08/20/2022. The subset consists of data from 2,334 counties, covering 242 weeks, and 19 diseases, formulated as a three-dimensional tensor $\boldsymbol{I}$. The NTF method (with the rank equal to 8) outputs three low-dimensional matrices, representing counties ($\boldsymbol{A}_1$), weeks ($\boldsymbol{A}_2$), and diseases ($\boldsymbol{D}$), and we show the mathematical details and an ablation study of rank selection in Supplementa. **Figure 1**(b) shows the output $\boldsymbol{D}$, which encodes the sparse representations of each pediatric disease and will be utilized as part of the features in the later prediction model.

We further apply the agglomerative clustering[29] with Ward algorithms[30] on the disease representations $\boldsymbol{D}$, resulting in a cluster hierarchy in **Figure 1**(c).

[3] https://mesonet.agron.iastate.edu/COOP/extremes.php

[4] https://mesonet.agron.iastate.edu/geojson/cli.py?dl=1&fmt=csv&dt={cur_date}

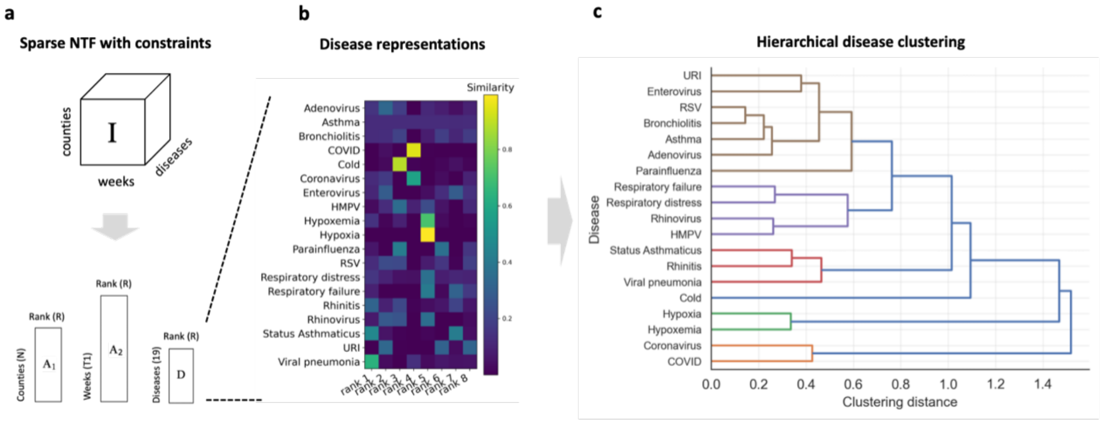


Figure 1. Disease hierarchical clustering by NTF. **a**. We formulated the medical claims data as a county-by-week-by-disease tensor (size: $2334 \times 242 \times 19$) and implemented the sparse non-negative tensor factorization (NTF) approach. The method outputs three low-dimensional matrices, representing counties ($\boldsymbol{A}_1$), weeks ($\boldsymbol{A}_2$), and diseases ($\boldsymbol{D}$); **b.** The matrix $\boldsymbol{D}$ encodes the sparse representations of each pediatric disease (adenovirus, asthma, bronchiolitis, etc.) as rows, with each disease has a sparse 8-dimensional similarity vector that indicates the similarity to one of the eight progression patterns (8 is the set rank of the NTF approach); **c.** We apply the agglomerative clustering algorithm with the Ward variance on the disease representations as the clustering metric, resulting in a disease cluster hierarchy. The x-axis is the clustering distance, and the y-axis shows the disease cluster hierarchy. Diseases closer in the hierarchy exhibit more similar trends across location and time.

RSV location clustering

Our next task is to track the RSV spread by analyzing its progression patterns from a geographic perspective. By identifying clusters of RSV cases in specific areas, healthcare providers can identify high-risk areas for potential outbreaks and take proactive steps for resource management in affected communities, such as appropriate stockpiling of medical services and supplies.

We analyze the RSV location clusters based on the medical claims (only the RSV portion up to 08/20/2022 before the prediction window) and Google search trend data independently. We first convert both data sources into location-by-week matrices ($2{,}334 \times 242$ for the claims and 51×261 for search data). We apply the same sparse non-negative tensor factorization (NTF) approach (due to space limitation, details in Supplementary Analysis 2) to obtain their location representation matrix $\boldsymbol{L}$. Each location becomes a 3-dimensional embedding vector. For medical claims data, the location matrix (of size $2{,}334 \times 3$, denoted as $\boldsymbol{L}$**)** represents the embeddings of each county (i.e., $\boldsymbol{l}_j$), which will be used in the later prediction model. For the Google trend data, the matrix $\boldsymbol{L}$ is of size 51×3 and represents the embedding of 51 states. The county-level claims data can provide more granular clustering results, and two datasets are analyzed individually.

In our study, we set rank 3 for NTF method (3 clusters). We observe three consistent RSV location clusters from both datasets, and the results are shown in **Figure 2**. Finer granular clustering results with more clusters could be achieved by increasing the size of tensor rank (i.e., number of final clusters). Further analysis on finding the best number of clusters is out of the scope.

Spatio correlation between location clusters

To provide a quantitative explanation for the clustering effect, we have also analyzed the correlation between the RSV outbreak date and the central longitude and latitude of each state over four RSV seasons (2018-2019, 2019-2020, 2021-2022, 2022-2023), using medical claims. RSV case counts were very low during the 2020-2021 season, presumably due to pandemic-related quarantine policies. We conducted Pearson correlation testing between the linear combination of longitude and latitude and the outbreak date of the state. The linear coefficient was chosen based on the best linear regression fit for each season. We found that longitude negatively correlates with the outbreak date while latitude positively correlates to the outbreak date, as shown in **Figure 3(a).**

Correlation between location clusters and climate factors

We further study the climate effect of the RSV location clusters. Previous studies[6,9-13] explored the relationship between the onset timing of RSV and the average precipitation and temperature, with some reporting a correlation of $R^2 \approx 0.5$ for both factors[9]. However, these studies had limitations, as they did not cover many states in the Midwest and North, and were conducted before 2013, which may not reflect recent trends, particularly after the COVID pandemic.

To address these gaps, we analyzed the last five years of RSV data from a large real-world medical claim database encompassing 2,334 counties and 51 states. We also utilized climatology data from the Iowa Environmental Mesonet project. Note that we defined the outbreak date as the day with the highest case count in each season, as the exact onset date can be difficult to pinpoint. Additionally, we used the average values from observatory stations located in each state to represent the annual average precipitation and temperature of the state. We found these two climate factors are highly correlated to RSV outbreak date (especially precipitation, with an $R^2 \approx 0.5$), shown in **Figure 3(b)(c)**.

Spatio-temporal RSV prediction

Having examined the spatio-temporal RSV patterns, we next present the predictive value of these patterns in estimating future RSV case counts for each county. In the prediction phase, *our goal is to accurately predict the recent pediatric RSV trend over the 2022-2023 winter season using the medical claims data before 08/20/2022 as observations.*

The predictions are based on sliding windows. During each prediction, we fix feature (i)(ii)(iii)(iv) and use the nearest Q weeks' claims data from feature (v) to predict the next S weeks' RSV trends. By default, we set the observation window $Q = 120$ (including the past one or two RSV seasons), and the prediction window $S = 4$ weeks (4 values predicted for each county).

We use the following information to extract the features: (i) disease representation $\boldsymbol{D} \in \mathbb{R}^{19\times R_1}$ (each disease has a representation $\boldsymbol{d}_i \in \mathbb{R}^{R_1}$) from disease clustering; (ii) the RSV geo-spatial embeddings $\boldsymbol{L} \in \mathbb{R}^{N\times R_2}$ (each county has a representation $\boldsymbol{l}_j \in \mathbb{R}^{R_2}$) from location clustering; (iii) static features: distance matrix $\boldsymbol{M}_1 \in \mathbb{R}^{N\times N}$, the mobility distance matrix $\boldsymbol{M}_2 \in \mathbb{R}^{N\times N}$, demographics statistics $\boldsymbol{M}_3 \in \mathbb{R}^{N\times 14}$, and the overall COVID, hospitalization and vaccination statistics $\boldsymbol{M}_4 \in \mathbb{R}^{N\times 39}$; (iv) the target timing, represented by month $m \in [1, 2, \dots, 12]$, and (v) the disease history time-series $\boldsymbol{H} \in \mathbb{R}^{N\times 120\times 19}$ (N counties, 120-week observation window and 19 diseases, which is a submatrix of claim data $\boldsymbol{I}$), to accurately predict the future trends of RSV at the county level. We show how to predict the target in the next $S = 4$ weeks for county j below. We use

$d = 32$ as our hidden dimension in the TAP-RSV model.

Feature 1: temporal disease embedding

We encode the disease similarity information and the trend of other diseases as the first feature. To obtain this, we consider the self-attention technique[37]. Upon the disease embedding matrix $\boldsymbol{D} \in \mathbb{R}^{19\times R_1}$ **,** we apply a linear layer $\boldsymbol{K}(\cdot): \mathbb{R}^{R_1} \rightarrow \mathbb{R}^d$ to generate the attention key, and a linear layer $\boldsymbol{Q}(\cdot): \mathbb{R}^{R_1} \rightarrow \mathbb{R}^d$ to generate the attention query. The self-attention matrix is obtained by cross product and an additional softmax layer.

$$\boldsymbol{A} = softmax(\boldsymbol{K}(\boldsymbol{D}) \cdot \boldsymbol{Q}(\boldsymbol{D})') \in \mathbb{R}^{19\times 19} \tag{21}$$

Here, $\boldsymbol{K}(\boldsymbol{D}), \boldsymbol{Q}(\boldsymbol{D}) \in \mathbb{R}^{19\times d}$ and $\boldsymbol{K}(\boldsymbol{D})\boldsymbol{Q}(\boldsymbol{D})'$ means the matrix cross product between $\boldsymbol{K}(\boldsymbol{D})$ and the transposition of $\boldsymbol{Q}(\boldsymbol{D})$**.** The weekly claims count of county j is $\boldsymbol{H}(j, :, :) \in \mathbb{R}^{120\times 19}$ (i.e., 120 weeks timeseries for 19 diseases). We apply the self-attention matrix to encode the interaction trends of 19 diseases in the *interaction embedding*.

$$\boldsymbol{E}(j, :, :) = \boldsymbol{H}(j, :, :)\boldsymbol{A} \in \mathbb{R}^{120\times 19} \tag{22}$$

We combine the original weekly timeseries and the interaction embedding into $\boldsymbol{H}'$.

$$\boldsymbol{H}'(j, :, :) = [\boldsymbol{H}(j, :, :), \boldsymbol{E}(j, :, :)] \in \mathbb{R}^{120\times(19+19)} \tag{23}$$

A recent work[31] applied the gated recurrent unit (GRU)[32], an effective type of recurrent neural network (RNN), to encode the longitudinal patient embedding. Inspired by this, we apply the GRU network with 38-dimensional input units and d-dim output units on $\boldsymbol{H}'$ to encode the temporal disease information and take the last output. The GRU network runs recurrently for 120 steps.

$$\boldsymbol{e}_{disease,j} = GRU\big(\boldsymbol{H}'(j, :, :)\big) \in \mathbb{R}^d \tag{24}$$

Feature 2 & 3: location and timing embedding

By referring to **Table** , we know that different locations can have distinct RSV patterns and timing is also a decisive factor in estimating the outbreak volume of a county. According to CDC[4], October to April are the typical RSV seasons in the north hemisphere. The target volume may be uniquely identified by the location clusters and the timing within a year.

We use the county location representation $\boldsymbol{l}_j \in \mathbb{R}^{R_2}$ from NTF results in Analysis 2 and apply a linear layer $\boldsymbol{h}(\cdot): \mathbb{R}^{R_2} \rightarrow \mathbb{R}^d$ to obtain the transformed location embedding.

$$\boldsymbol{e}_{location,j} = \boldsymbol{h}(\boldsymbol{l}_j) \in \mathbb{R}^d \tag{25}$$

We learn the month representation $\boldsymbol{M} \in \mathbb{R}^{12\times d}$ over 12 months to indicate the timing and use the month embedding of the prediction window to encode the month information

$$\boldsymbol{e}_{timing,j} = \boldsymbol{M}(m) \in \mathbb{R}^d \tag{26}$$

Here, $m \in [1, 2, \dots, 12]$ is the target month index.

Feature 4: static feature embedding

Demographics and other static healthcare features are also great descriptive features for RSV prediction. For example, affluent neighborhood may have more hospital resources and are more willing to invest on clinical supply. Thus, their RSV rate might be low or decrease quickly. To preprocess the static features, we normalize them by subtracting out the mean signal and scaling by standard deviations. After processing, the numerical values of the same feature across different counties are normal distributed. We then concatenate the static features and devise a linear prediction layer $\boldsymbol{s}(\cdot): \mathbb{R}^{N+N+14+39} \rightarrow \mathbb{R}^d$ for learning hidden embeddings.

$$\boldsymbol{e}_{static,j} = \boldsymbol{s}([\boldsymbol{M}_1(j, :), \boldsymbol{M}_2(j, :), \boldsymbol{M}_3(j, :), \boldsymbol{M}_4(j, :)]) \in \mathbb{R}^d \tag{27}$$

Multi-faceted non-negative prediction

In the last layer, we concatenate the previous four feature embeddings and apply a Gaussian Error Linear Unit (GELU)[33] as the activation and design a final prediction network by a two-layer neural network $\boldsymbol{f}(\cdot): \mathbb{R}^{4d} \rightarrow \mathbb{R}^{S}$ ($4d$ is for the concatenation of 4 different feature embeddings and $S = 4$ is the number of future weeks in prediction). We further apply an exponential function to prevent the negative values. The final objective is MSE loss, measuring the gap between the prediction and true values.

$$\widehat{\boldsymbol{y}}_j = \exp \boldsymbol{f}\left(GELU\left(\left[\boldsymbol{e}_{disease,j}, \boldsymbol{e}_{location,j}, \boldsymbol{e}_{timing,j}, \boldsymbol{e}_{static,j}\right]\right)\right) \in \mathbb{R}^{S} \qquad (28)$$

Adding the exponential function also has statistical benefits. The real target counts usually follow a long-tail distribution, while their log-values are usually normal distributed. A recent work[46] considered this phenomenon and predicted the log-transformed counts instead of the real target. Our exponential function has a similar effect, which encourages the output of function $\boldsymbol{f}$ to follow a normal distribution and improves the training performance.

Results

Analysis 1: Disease clustering hierarchy

Figure 1(c) shows the obtained disease clustering hierarchy. We observe that the results align with clinical knowledge, as ontologically and pathophysiologically related diseases are grouped together. Specifically, we note the following examples. Enterovirus is a common cause of URI and causes bronchiolitis. Virtually all cases of bronchiolitis begin as URI. Parainfluenza and adenovirus are common viral causes of bronchiolitis, and infection with such viruses is a common cause of acute asthma exacerbation. Respiratory failure and respiratory distress and considered degrees of severity of one another. Status asthmaticus, rhinitis, and viral pneumonia are much more commonly diagnosed in children over the age of 2 years (in contrast to bronchiolitis, which occurs by definition only in children under 2 years). Hypoxia and hypoxemia are physiologically distinct but often used interchangeably in clinical documentation. COVID-19 is caused by a variant form of coronavirus. This hierarchy allows us to gain a quantitative understanding of the relationship among these diseases.

Analysis 2: RSV location clusters

Three consistent location clusters from medical claims and Google search data

Figure 2(a) displays the learned progression patterns and their corresponding county clusters extracted from the claim data. The color intensity in the map is obtained from the NTF model. A darker color represents a higher similarity between the county's trend and one of three NTF-extracted trend patterns. We found that RSV always appears first in cluster 1, followed by cluster 2, and then cluster 3, over each season. Independently, we train another NTF model and extract state clusters from the Google data in **Figure 2(b),** and its trend patterns are shown in **Figure 3 (supp)** in **Supplementa** materials. Interestingly, we observe a significant alignment in the county-level and state-level clusters. The algorithm identifies three clusters with obvious geographical similarities – Southeast, Central, and Northwest regions.

Temporal shift observed in CDC surveillance data of 12 states

In **Figure 2(c),** we further line up the actual surveillance trends on the timeline roughly following the cluster orders. Georgia and Tennessee are in the first cluster (Southeast region); Colorado, Connecticut, Maryland, Michigan, Minnesota, New Mexico, and New York are in the second cluster (Central, Northeast, and Southwest regions); Utah, California, Oregon are in the third cluster (West and Northwest regions). We observe a clear outbreak date shift across areas from East to West and from South to North.

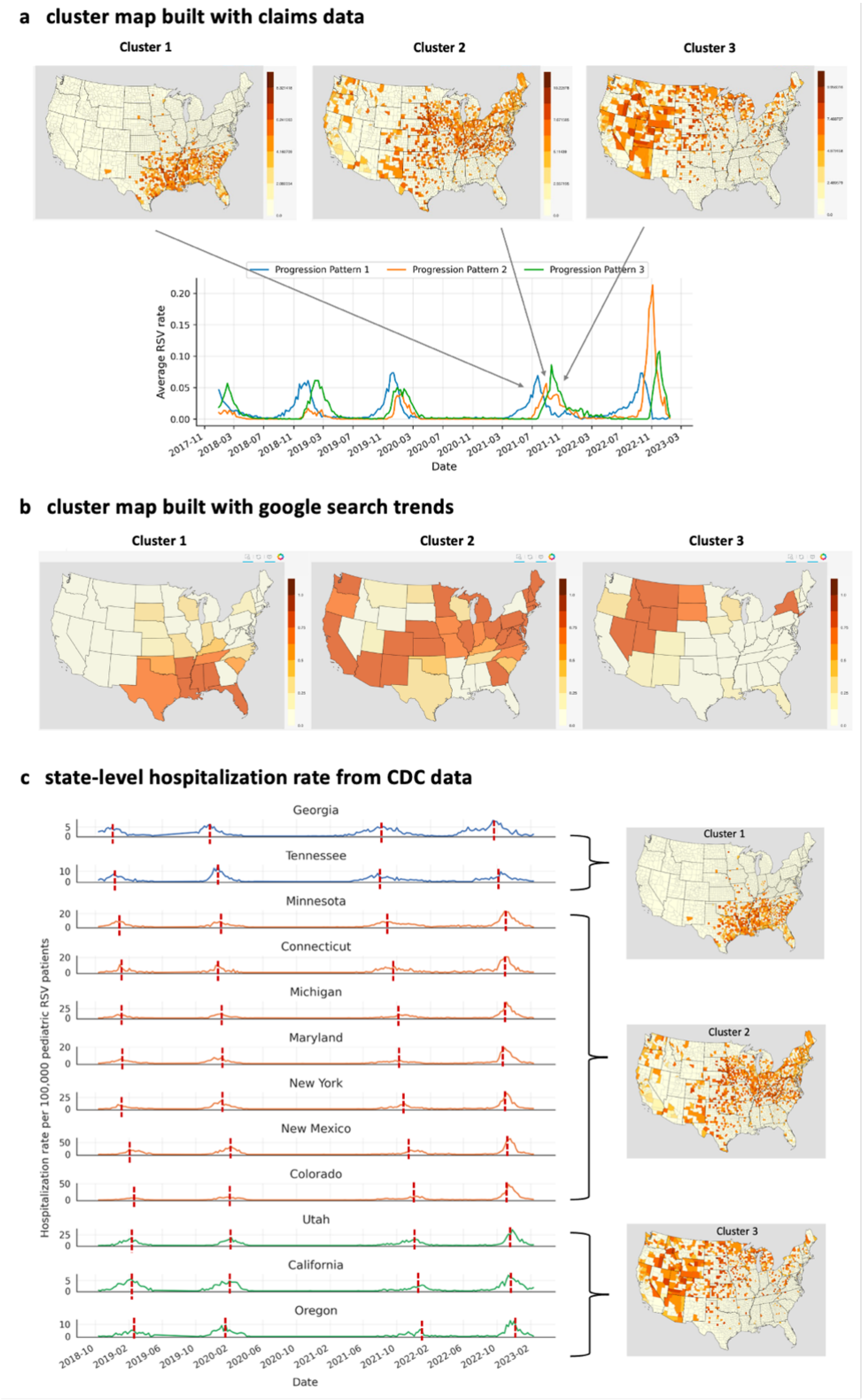


Figure 2. Three consistent RSV location clusters from different sources. **a.** We applied the sparse NTF

approach to the county-by-week claims data (size: 2334×242) on the RSV portion and obtained three location clusters, and the temporal progression patterns. A darker color indicates higher similarity to the clustering center. The progression patterns are shared in the same cluster, and we plotted the three clusters in order. The disease progression of cluster 1 occurred one to two weeks earlier than cluster 2, and three to four weeks earlier than cluster 3 over each season; **b.** The same NTF approach was independently applied to the Google search trend data. The extracted three clusters were similar at the state level; **c.** For the CDC surveillance data, we match the 12 states to three location clusters extracted from medical claim data. The red dotted line indicates the peak of each RSV hospitalization season. We observed a clear temporal shift of RSV curves from the top to the bottom (from cluster 1 to cluster 3) along the timeline. This figure independently confirms three distinct RSV patterns across the Southeast, Central, and Northwest regions.

Correlation between shift and longitude, latitude, precipitation, and temperature

Figure 3(a) shows the linear combination of longitude and latitude and the RSV outbreak date of the state. Our analysis reveals that longitude negatively correlates with the outbreak date, indicating that eastern regions experience outbreaks earlier than western regions. Conversely, latitudes positively correlate with the outbreak, suggesting that southern states are more likely to experience outbreaks earlier than northern regions. The best-fitting curves correlate highly with the outbreak dates (with an $R^2 \approx 0.4$ for the last three seasons).

Figure 3(b) shows the correlation between outbreak date and annual precipitation at state level. From the figures, we can observe that the correlation remains relatively stable during each season, with correlation values $R^2 \approx 0.5$, and a higher volume of precipitation leads to an earlier timing of the outbreak. However, the correlation in the first and last season ($R^2 > 0.5$) is higher than in the middle two seasons ($R^2 < 0.5$), which could be due to the social distancing policies of the COVID pandemic[16].

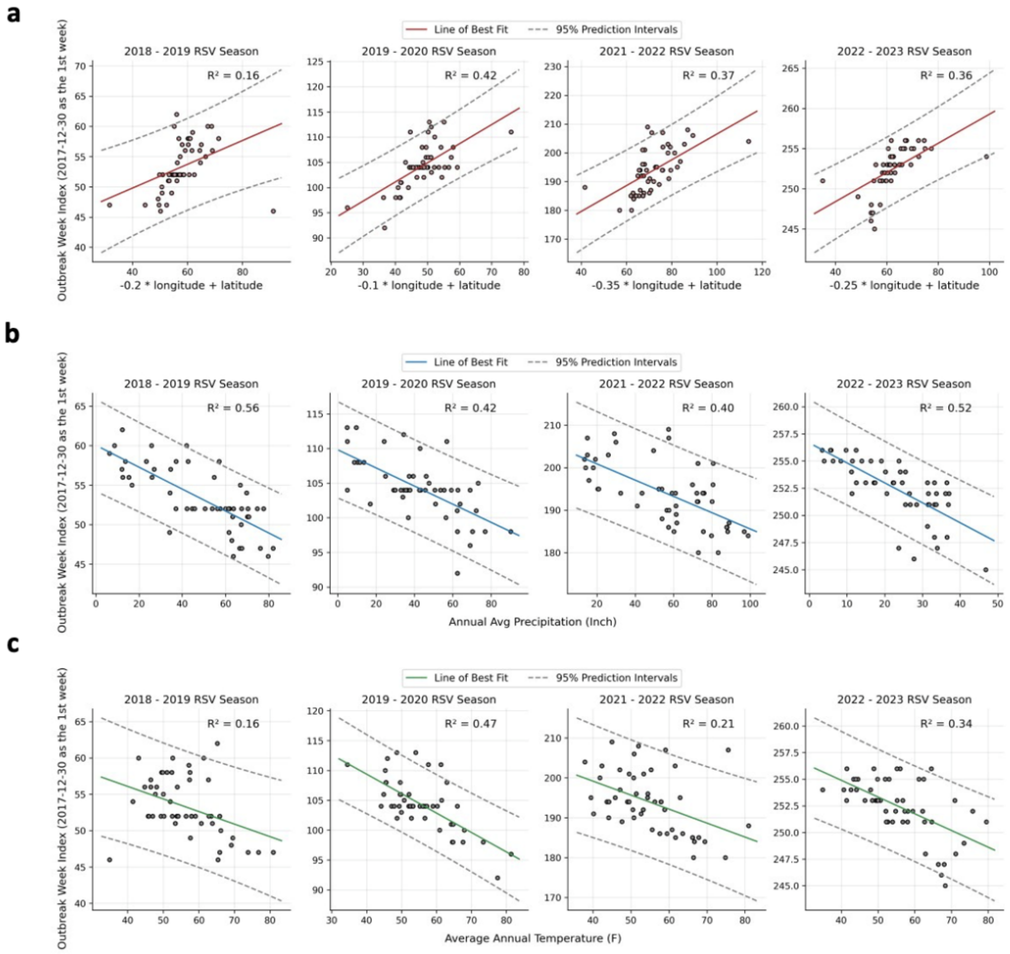


Figure 3. Correlation to latitude & longitude, precipitation, and temperature. **a.** We test the correlation between RSV and the best linear combination of raw longitude and latitude value and breakdown for four RSV seasons. The y-axis represents the outbreak week index (12/20/2017 as the first week in index), and x-axis represents the best linear combination of raw values of longitude and latitude. Each dot (in total 51) in the figures represents a state in the US. The outbreak date has regional differences with a correlation score around $R^2 \approx 0.4$ for the last three seasons. We also find that the longitude is generally negatively correlated to the outbreak date, meaning that eastern regions will outbreak earlier than western regions. The latitude is positively correlated to the outbreak dates, meaning that southern regions outbreaks earlier than northern regions; **b.** Similar linear correlation test has been conducted between RSV outbreak dates and annual precipitation of a state. We find precipitation has an average correlation $R^2 \approx 0.5$ to RSV. The relatively low correlation in 2019-2020 and 2021-2022 season might be affected by the global COVID pandemic; **c.** Correlation test between RSV outbreak dates and annual average temperature of a state. The correlation is not stable across the four seasons, while the 2019-2020 season has a high score $R^2 \approx 0.47$. By comparison, the temperature factor has a weaker correlation to RSV outbreak than precipitation.

Figure 3(c) shows the correlation to annual temperature, which roughly indicates that higher temperatures result in earlier outbreak dates. This can be partially explained by the fact that solar[34] and high temperature[35] could inactivate many microbes including RSV. However, the correlation is shown to be unstable across different seasons. For example, 2019-2020 season gives a $R^2 = 0.47$ score while the other three seasons demonstrate a weak correlation of $R^2 < 0.34$. Comparing the two factors, we conclude that precipitation has a stronger correlation to RSV outbreak than temperature.

Prediction: Multi-faceted RSV Trend Forecasting

In the prediction phase, *our goal is to accurately predict the recent pediatric RSV trend over the 2022-2023 winter season using the medical claims data before 08/20/2022 as observations.*

Experimental Settings

Note that the 2022-2023 RSV season began earlier than in previous years and spans from 08/20/2022 to 12/09/2022, based on the medical claims. In the experiment, the season is divided into four prediction windows: 08/20/2022 - 09/17/2022 - 10/15/2022 - 11/12/2022 - 12/09/2022. Our prediction model is initially trained on the 120 training weeks up to 7/22/2022 and the best hyperparameters are selected using the 120 validation weeks up to 8/20/2022. The model is then evaluated on the first prediction window (from 08/20/2022 to 09/17/2022). Next, the training and validation windows are shifted forward by 4 weeks, and the model is retrained and the best hyperparameters are reselected to evaluate on the second prediction window (from 09/17/2022 to 10/15/2022). This process is repeated until all four prediction windows are evaluated. An illustration can be found in **Figure 4(a)**. Four predictions are conducted independently.

To evaluate the result of prediction, we use the mean square error (MSE), mean absolute error (MAE), the coefficient of determination (R^2), and the Pearson correlation coefficient (PCC) to assess the prediction performance. The range of MSE and MAE are $(0, +\infty)$, lower values are better. The range of R^2 is $(-\infty, 1)$, and the range of PCC is $(-1, 1)$, higher values are better.

We implement the following baseline methods from different perspectives:

- **ARIMA** is a popular time series analysis model used for forecasting future values based on past observations.
- **LSTM**[36] is a type of RNN architecture that is used for sequence prediction. The hidden dimension is set to 32.
- **Transformer**[37] is popular for long-range information preserving in sequence modeling. We set the number of attention heads to 4, the number of attention layers to 2, and the hidden dimension to 32.
- **XGBoost**[38] is commonly used for regression tasks on tabular features. In this study, we extract the RSV volume, and the first, second, and third-order statistics of the trend as the hand-crafted time-series features, combined with static features for prediction.
- **STAN**[39] is an RNN-based spatio-temporal prediction model, initially developed for COVID-19 case prediction. We adopt it for RSV prediction. This model also integrates the spatial map graph neural network module. We set the graph network dimension to 64, and the hidden dimension to 32. The SIR dynamics module of STAN does not apply to our prediction task, and we thus remove the SIR constraints in the final objective function.
- **HOIST**[25] is a recent spatio-temporal model for COVID-19 case prediction with the Ising dynamics constraints. We adopt it for spatio-temporal RSV prediction. The hidden dimension of GRU module is set to 32 and the graph neural network embedding is set to

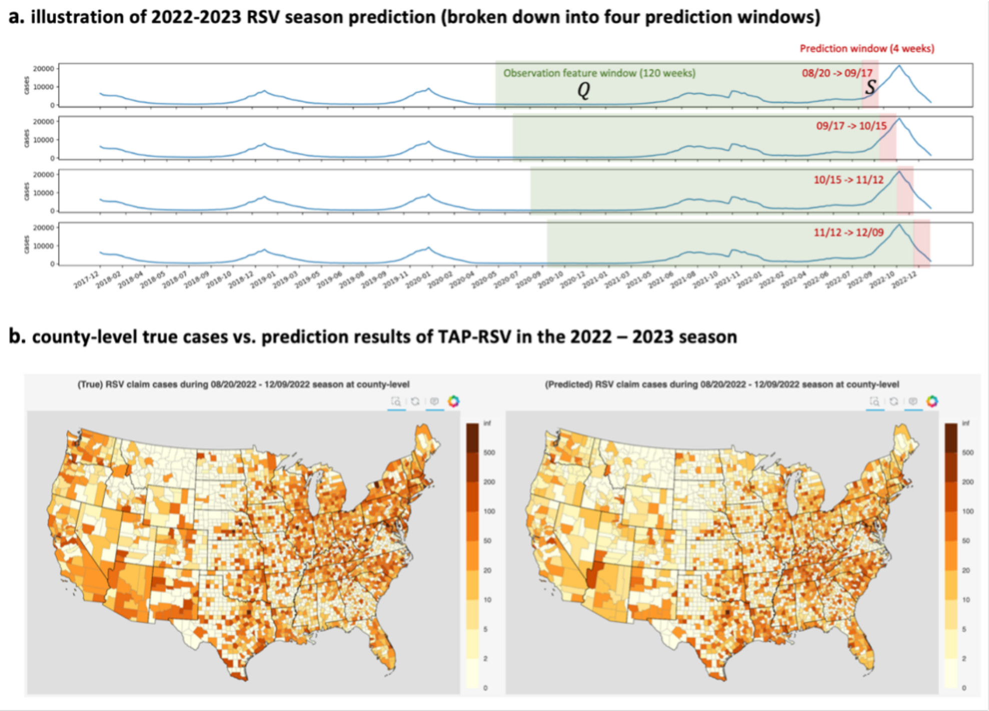
64, the same hyperparameters as STAN.

All models are trained five times with different random seeds, and the mean and standard deviation values are reported. **The source code of all models is publicly available on GitHub**[5]. We have also provided the feature sets of each model in Supplementary (**Table 4** (supp)).

Better performance in 2022-2023 RSV season prediction

In **Figure 4(a)**, we illustrate the main prediction setting for our study. We divide the four-month RSV period into four prediction windows (4 weeks each) and use the previous 120 weeks as the observation. **Table 1** shows the comparison of all models over four prediction windows. Our TAP-RSV model performs best on almost all metrics except for the PCC value against HOIST on the first prediction window. We report the test p-values of **Table 1** under five random seeds in Supplementa (**Table 4**), which shows that our performance gain is significant in most cases with p-values <0.05.

We find that the naïve ARIMA model gives the worst performance, with PCC values < 0.02 on all prediction windows. This is because ARIMA only supports the raw time-series and cannot utilize static features. We also find that the LSTM and Transformer baselines, which concatenate the static features into each time step, are relatively strong among all baselines. STAN and HOIST are two spatio-temporal models that have specially designed spatial graph convolutions initially for pandemic prediction. However, the performance of STAN is not desirable when adopted for our task, probably due to the differences in problem settings. Among all baselines, HOIST shows the best performance consistently. The poor performance of XGBoost model implies that modeling the raw sequences like in Transformer and LSTM is more powerful than using time-series statistical features.


a. illustration of 2022-2023 RSV season prediction (broken down into four prediction windows)

b. county-level true cases vs. prediction results of TAP-RSV in the 2022 – 2023 season

Figure 4. Illustration of 2022-2023 RSV season prediction. **a.** Broken down 2022-2023 RSV season

[5] https://github.com/ycq091044/TAP-RSV

prediction into four 4-week windows. In each window, we use the extracted features from the nearest 120 weeks (long enough to include the past one or two RSV seasons) as prediction features to forecast the next four weeks at the county-level. Each window is shifted by four weeks. Q means the feature window length and S means the prediction window length (unit: week); **b.** County-level true cases vs. prediction results of TAP-RSV in the 2022 – 2023 season. The left figure shows the true total volume of each county, and the right figure shows the total predicted volume. For both figures, we sum up the cases over the entire 2022 – 2023 season.

To give a visual illustration, we further compare the true county-level RSV volume in the entire 2022-2023 RSV season with the predicted volume over the season in **Figure 4(b)**. We observe that the volume distributions across all counties are very similar in the two figures, but the true figure generally looks darker, indicating that it has a larger scale in volume. We show that by plotting the distribution of true and predicted case counts in Supplementa (**Figure 6 (supp)**). The scale discrepancy can be explained by the unprecedented case surge in the current RSV season, and the extracted prediction features are not strong enough to capture the unexpected volume.

Table 1. Model performance comparison on 2022-2023 four prediction windows. We bold-font the best metric and shade our model outputs. Format of the table: mean value ± standard deviations over 5 random seeds. Our TAP-RSV model gives the best performance on most of the metrics with $p < 0.05$, except the PCC value against HOIST on the first prediction window. We show the p-values in **Table** in Supplementa.

Window	Model	MSE	MAE	PCC	$\boldsymbol{R^2}$
08/20 ➔ 09/17	ARIMA	36.11 ± 1.3054	3.33 ± 0.0231	0.01 ± 0.0123	-0.56 ± 0.1341
	LSTM	22.02 ± 0.3721	2.45 ± 0.0438	0.77 ± 0.0081	0.35 ± 0.0113
	Transformer	22.90 ± 0.7879	2.46 ± 0.0463	0.74 ± 0.0127	0.32 ± 0.0232
	XGBoost	31.96 ± 0.0126	2.47 ± 0.0013	0.62 ± 0.0004	0.08 ± 0.0004
	STAN	35.74 ± 1.2799	2.68 ± 0.0801	0.48 ± 0.0180	-0.06 ± 0.0362
	HOIST	20.01 ± 1.9669	2.36 ± 0.0274	**0.79 ± 0.0121**	0.40 ± 0.0466
	TAP-RSV	**19.26 ± 1.3043**	**2.29 ± 0.0859**	0.78 ± 0.0102	**0.41 ± 0.0243**
09/17 ➔ 10/15	ARIMA	45.45 ± 1.1453	3.60 ± 0.0643	0.01 ± 0.0134	-0.53 ± 0.0743
	LSTM	59.13 ± 3.2203	3.69 ± 0.1173	0.75 ± 0.0143	0.25 ± 0.0435
	Transformer	59.22 ± 1.9690	3.64 ± 0.0808	0.75 ± 0.0119	0.25 ± 0.0260
	XGBoost	82.43 ± 0.0001	4.09 ± 0.0000	0.71 ± 0.0000	-0.02 ± 0.0000
	STAN	94.88 ± 0.8759	4.47 ± 0.0390	0.50 ± 0.0176	-0.20 ± 0.0116
	HOIST	48.68 ± 5.7941	3.50 ± 0.0774	0.77 ± 0.0277	0.37 ± 0.0615
	TAP-RSV	**41.66 ± 4.7963**	**3.26 ± 0.0722**	**0.79 ± 0.0079**	**0.46 ± 0.0494**
10/15 ➔ 11/12	ARIMA	62.94 ± 2.1096	4.74 ± 0.0589	0.02 ± 0.0432	-0.49 ± 0.0975
	LSTM	75.97 ± 3.5810	4.10 ± 0.0450	0.68 ± 0.0058	0.32 ± 0.0311
	Transformer	80.78 ± 4.4693	4.29 ± 0.0684	0.69 ± 0.0049	0.28 ± 0.0417
	XGBoost	121.8 ± 0.0104	5.38 ± 0.0004	0.71 ± 0.0004	-0.08 ± 0.0001
	STAN	138.4 ± 1.7601	5.67 ± 0.0464	0.48 ± 0.0343	-0.25 ± 0.0166
	HOIST	63.61 ± 6.8832	4.11 ± 0.0534	0.74 ± 0.0130	0.41 ± 0.0545
	TAP-RSV	**53.30 ± 5.4441**	**3.99 ± 0.0625**	**0.75 ± 0.0169**	**0.50 ± 0.0392**

11/12 ➔ 12/09	ARIMA	83.69 ± 3.9015	3.78 ± 0.1345	0.02 ± 0.0341	-2.23 ± 0.0132
	LSTM	38.34 ± 1.2719	3.38 ± 0.0687	0.69 ± 0.0122	0.45 ± 0.0139
	Transformer	40.25 ± 1.6947	3.32 ± 0.0759	0.69 ± 0.0100	0.40 ± 0.0320
	XGBoost	70.25 ± 0.0165	4.25 ± 0.0004	0.69 ± 0.0003	0.02 ± 0.0002
	STAN	84.57 ± 3.5291	4.49 ± 0.2036	0.43 ± 0.0443	-0.23 ± 0.0557
	HOIST	33.77 ± 5.7051	2.97 ± 0.0364	0.73 ± 0.0401	0.49 ± 0.0698
	TAP-RSV	**29.80 ± 2.3285**	**3.09 ± 0.1531**	**0.75 ± 0.0163**	**0.52 ± 0.0463**

Discussion

Summary: Our work introduces a comprehensive Tensor-based Analysis and Prediction (TAP) framework for studying the respiratory syncytial virus (RSV), utilizing multiple data sources. Specifically, we incorporate five-year county-level pediatric claims data with 19 pediatric diseases, state-level CDC surveillance data (pediatric portion) from 12 participant states, state-level Google RSV keyword search trends from 51 states, five-year climate observation data from Iowa State University covering 51 states, and county-level static features related to demographics, mobility distances, vaccination, hospitalization, and COVID statistics for 2334 counties.

We use the sparse non-negative tensor factorization (NTF) method and extract a clinically meaningful disease hierarchy with quantitative embeddings. We also analyze RSV disease's location distribution and find meaningful RSV location patterns over the US. The Southeast region of the US tends to have the RSV peak earlier than other regions of the US. The Central and Northeast regions will follow up and peak one or two weeks later, while the West and Midwest regions are always the last to peak. The clustering results are consistent across all three data sources. Further, we find that the annual precipitations and temperatures are two potential correlative factors (with average $R^2 \approx 0.5$ and $R^2 \approx 0.3$ via linear correlation test) explaining the peak shift. Our finding is drawn from RSV data from the last five years, extending from previous works[6,9] (using data up to 2013). We predict the recent 2022 – 2023 RSV season with multi-faceted features, which gives decent performance in the 2022-2023 RSV season (MSE < 55, MAE < 4, PCC > 0.75, and $R^2 \approx 0.5$) at the county level. Our TAP-RSV model performs better than all baselines with $p < 0.05$ in most cases.

Limitations: Our research has some limitations that should be acknowledged. Firstly, in data curation, we borrowed demographic information such as pediatric population, race ratio, and income level from a previous study[24], which collected data from different sources over several years. Secondly, we used claims data as our primary resource for predicting RSV severity, which may underestimate the true number of RSV cases. During the correlation test, we approximated state-level climate statistics using average values from different observatory stations, which may introduce inaccuracies. Despite these limitations, our study results provide valuable insights, and our model can be applied to other high-quality RSV data if available.

For modeling, our methodology involves extracting different feature sets separately and then concatenating them into a final prediction layer. This approach has already yielded promising results, outperforming strong baseline models. However, further improvements could be achieved by incorporating feature crossing or utilizing advanced techniques, such as graph neural networks.

Take-aways and future works: Our proposed method can inspire follow-up works in the following ways. First, the main techniques mentioned in the paper – non-negative tensor factorization (NTF)

– can be utilized in many other applications to extract the common patterns/clusters (especially for location clusters or population groups) from high dimensional data sources. Also, the NTF methods are specifically suitable for handling problems with limited data sources compared to using deep learning models. Second, this paper uses other diseases time series to improve the prediction of RSV trend, and the same techniques can be used for the analysis or prediction of target by using rich information from other relevant variables. Additionally, combining Google search data as a strong supplementary for early-stage prediction can be a promising direction in similar applications since this data source is easy to access and more recent data can be available as the timely prediction features. However, Google search data should also be treated carefully before used in critical healthcare scenarios, as it does not directly reflect health information of individuals or a population.

Data Availability Statement

The CDC surveillance data is freely collected from RSV-Net https://www.cdc.gov/rsv/research/rsv-net/dashboard.html. The Google search data is freely collected from Google trend API https://trends.google.com/trends/explore. The climatology data is freely collected from Iowa State University Environmental Mesonet project https://mesonet.agron.iastate.edu/COOP/extremes.php. The medical claims data are proprietary that contain sensitive healthcare information and are extracted from https://www.iqvia.com/solutions/real-world-evidence/ but can be accessed on request. The demographics data is freely downloaded from https://github.com/JieYingWu/COVID-19_US_County-level_Summaries. Other public data are available at https://github.com/ycq091044/TAP-RSV. All the source data of the figures in the manuscript are uploaded to https://github.com/ycq091044/TAP-RSV/blob/main/Source-Data.xlsx.

Code Availability Statement

The codes for baseline and model construction, training and inference used in this paper are publicly available at https://github.com/ycq091044/TAP-RSV.

Competing Interest Statement

The authors declare that there are no competing interests.

Authors Statement

CY and JG processed the feature and built the model. JS and AC provided technical and clinical guidance. LG provided the medical claims dataset. CY, JG, AC and JS participated in report writing. All authors declare that they have no conflicts of interest. All correspondence can be sent to jimeng.sun@gmail.com.

Acknowledgement

We thank Dr. Mustafa Bakir from OSF Healthcare for assisting with the literature search.

Funding Statement

This work was supported by NSF award SCH-2205289, SCH-2014438, IIS-1838042, NIH award R01 1R01NS107291-01.

Inclusion & Ethics Statement

The study was approved by the University of Illinois Institute Review Board with the project title "Non-negative Tensor Factorization of Pediatric RSV Infections" and IRBNet ID 201174-1.

References

1. Suh, M.*, et al.* Respiratory syncytial virus is the leading cause of United States infant hospitalizations, 2009–2019: a study of the national (nationwide) inpatient sample. *The Journal of Infectious Diseases* **226**, S154-S163 (2022).
2. Hall, C.B.*, et al.* The burden of respiratory syncytial virus infection in young children. *New England Journal of Medicine* **360**, 588-598 (2009).
3. Suh, M., Movva, N., Bylsma, L.C., Fryzek, J.P. & Nelson, C.B. A systematic literature review of the burden of respiratory syncytial virus and health care utilization among United States infants younger than 1 year. *The Journal of Infectious Diseases* **226**, S195-S212 (2022).
4. Rainisch, G., Adhikari, B., Meltzer, M.I. & Langley, G. Estimating the impact of multiple immunization products on medically-attended respiratory syncytial virus (RSV) infections in infants. *Vaccine* **38**, 251-257 (2020).
5. Bennett, M.V., McLaurin, K., Ambrose, C. & Lee, H.C. Population-based trends and underlying risk factors for infant respiratory syncytial virus and bronchiolitis hospitalizations. *PLoS One* **13**, e0205399 (2018).
6. Obando-Pacheco, P.*, et al.* Respiratory syncytial virus seasonality: a global overview. *The Journal of infectious diseases* **217**, 1356-1364 (2018).
7. Yu, J.*, et al.* Respiratory syncytial virus seasonality, Beijing, China, 2007–2015. *Emerging infectious diseases* **25**, 1127 (2019).
8. Zheng, Z., Warren, J.L., Artin, I., Pitzer, V.E. & Weinberger, D.M. Relative timing of respiratory syncytial virus epidemics in summer 2021 across the United States was similar to a typical winter season. *Influenza and Other Respiratory Viruses* **16**, 617-620 (2022).
9. Baker, R.E.*, et al.* Epidemic dynamics of respiratory syncytial virus in current and future climates. *Nature communications* **10**, 5512 (2019).
10. Chan, P.*, et al.* Epidemiology of respiratory syncytial virus infection among paediatric patients in Hong Kong: seasonality and disease impact. *Epidemiology & Infection* **123**, 257-262 (1999).
11. Pitzer, V.E.*, et al.* Environmental drivers of the spatiotemporal dynamics of respiratory syncytial virus in the United States. *PLoS pathogens* **11**, e1004591 (2015).
12. Sloan, C.*, et al.* The impact of temperature and relative humidity on spatiotemporal patterns of infant bronchiolitis epidemics in the contiguous United States. *Health & place* **45**, 46-54 (2017).
13. Chew, F., Doraisingham, S., Ling, A., Kumarasinghe, G. & Lee, B. Seasonal trends of viral respiratory tract infections in the tropics. *Epidemiology & Infection* **121**, 121-128 (1998).
14. Paynter, S. Humidity and respiratory virus transmission in tropical and temperate settings. *Epidemiology & Infection* **143**, 1110-1118 (2015).
15. Foley, D.A.*, et al.* Examining the interseasonal resurgence of respiratory syncytial virus in Western Australia. *Archives of disease in childhood* **107**, e1-e7 (2022).
16. Hatter, L., Eathorne, A., Hills, T., Bruce, P. & Beasley, R. Respiratory syncytial virus: paying the immunity debt with interest. *The Lancet Child & Adolescent Health* **5**, e44-e45 (2021).
17. Peter J. Stein, P., DC. This Year's RSV Surge: Bigger, Earlier, and Affecting Older Patients Than Previous Seasonal Outbreaks. in *Clinical Advisor* (2022).
18. Reis, J., Yamana, T., Kandula, S. & Shaman, J. Superensemble forecast of respiratory syncytial virus

outbreaks at national, regional, and state levels in the United States. *Epidemics* **26**, 1-8 (2019).

19. Korsten, K.*, et al.* Prediction model of RSV-hospitalization in late preterm infants: an update and validation study. *Early Human Development* **95**, 35-40 (2016).
20. Simões, E.A.*, et al.* A predictive model for respiratory syncytial virus (RSV) hospitalisation of premature infants born at 33–35 weeks of gestational age, based on data from the Spanish FLIP Study. *Respiratory research* **9**, 1-10 (2008).
21. Gebremedhin, A.T., Hogan, A.B., Blyth, C.C., Glass, K. & Moore, H.C. Developing a prediction model to estimate the true burden of respiratory syncytial virus (RSV) in hospitalised children in Western Australia. *Scientific Reports* **12**, 1-12 (2022).
22. Prevention, C.f.D.C.a. RSV-NET: Respiratory Syncytial Virus Hospitalization Surveillance Network. (2021).
23. Kang, Y.*, et al.* Multiscale dynamic human mobility flow dataset in the US during the COVID-19 epidemic. *Scientific data* **7**, 390 (2020).
24. Killeen, B.D.*, et al.* A county-level dataset for informing the United States' response to COVID-19. *arXiv preprint arXiv:2004.00756* (2020).
25. Junyi Gao, J.H., Christina Mack, Lucas Glass, Adam Cross, Jimeng Sun. HOIST: Evidence-Driven Spatio-Temporal COVID-19 Hospitalization Prediction with Ising Dynamics. (2023).
26. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* **20**, 533-534 (2020).
27. Shashua, A. & Hazan, T. Non-negative tensor factorization with applications to statistics and computer vision. in *Proceedings of the 22nd international conference on Machine learning* 792-799 (2005).
28. Lee, D. & Seung, H.S. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* **13**(2000).
29. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378* (2011).
30. Szekely, G.J. & Rizzo, M.L. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of classification* **22**, 151-184 (2005).
31. Yang, C., Xiao, C., Ma, F., Glass, L. & Sun, J. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711* (2021).
32. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
33. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
34. Aune, K.T., Davis, M.F. & Smith, G.S. Extreme Precipitation Events and Infectious Disease Risk: A Scoping Review and Framework for Infectious Respiratory Viruses. *International journal of environmental research and public health* **19**, 165 (2021).
35. Sloan, C., Moore, M.L. & Hartert, T. Impact of pollution, climate, and sociodemographic factors on spatiotemporal dynamics of seasonal respiratory viruses. *Clinical and translational science* **4**, 48-54 (2011).
36. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735-1780 (1997).
37. Vaswani, A.*, et al.* Attention is all you need. *Advances in neural information processing systems* **30**(2017).

38. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* 785-794 (2016).

39. Gao, J.*, et al.* STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *Journal of the American Medical Informatics Association* **28**, 733-743 (2021).

40. Yang, C.*, et al.* Mtc: Multiresolution tensor completion from partial and coarse observations. in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 1953-1963 (2021).

41. Yang, C.*, et al.* ATD: Augmenting CP Tensor Decomposition by Self Supervision. *Advances in neural information processing systems* (2022).

42. Golub, G.H. & Von Matt, U. *Tikhonov regularization for large scale problems*, (Citeseer, 1997).

43. Chen, G. & Teboulle, M. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming* **64**, 81-101 (1994).

44. Parikh, N. & Boyd, S. Proximal algorithms. *Foundations and trends® in Optimization* **1**, 127-239 (2014).

45. Yang, C., Qian, C. & Sun, J. GOCPT: Generalized Online Canonical Polyadic Tensor Factorization and Completion. *arXiv preprint arXiv:2205.03749* (2022).

46. Gao, J.*, et al.* MedML: Fusing medical knowledge and machine learning models for early pediatric COVID-19 hospitalization and severity prediction. *Iscience* **25**, 104970 (2022).

Supplemental Materials

Organization of Supplemental materials: **First,** we describe the model details and implementation procedures of NTF in disease clustering and location clustering; **Then,** we list the static features used for prediction; **Next,** we show additional experimental results (1 – 7), including ablation studies and hypothetis testing, to support claims in the main text. Note that we follow the recommendations set out in the Global Code of Conduct for Research in Resource-Poor Settings when designing, executing, and reporting the research and this research does not use individual-level data which may raise ethic issues. Our study complies with the recommendations of the GATHER statement.

Analysis 1: Non-negative Tensor Factorization for Disease Clustering

To understand the progression of RSV diseases, this paper first explores its relationship with other 18 pediatric diseases. We use the claims data as the resource. Non-negative tensor factorization (NTF) [27,28] is commonly used for extracting the low-rank structure of real-world count-based data and finding dominant spatio-temporal patterns. In this task, we use the sparse NTF approach to learn the representations of 19 diseases $\boldsymbol{d}_i$ $(i = 1, 2, \ldots, 19)$.

Sparse non-negative tensor factorization (NTF)

Formally, given county-by-week-by-disease tensor based on claims data $\boldsymbol{I} \in \mathbb{R}^{N\times T_1\times 19}$ (we use the portion up to 08/20/2022 and still use $\boldsymbol{I}$ as the notation for convenience), the rank-R_1 NTF approach will decompose it into: a county representation matrix $\boldsymbol{A_1} \in \mathbb{R}^{N\times R_1}$, a week representation matrix $\boldsymbol{A}_2 \in \mathbb{R}^{T_1\times R_1}$, and a disease representation matrix $\boldsymbol{D} \in \mathbb{R}^{19\times R_1}$, following the canonical polyadic decomposition (CPD) model [40,41]. Here, we only care about the disease matrix $\boldsymbol{D}$, where each row is $\boldsymbol{d}_i \in \mathbb{R}^{R_1}$, and R_1 is called the rank number (interpreted as the representation dimension). We use $R_1 = 8$ in the experiments. An ablation study of R_1 is provided in Supplemental materials (**Fig. 2 (supp)**).

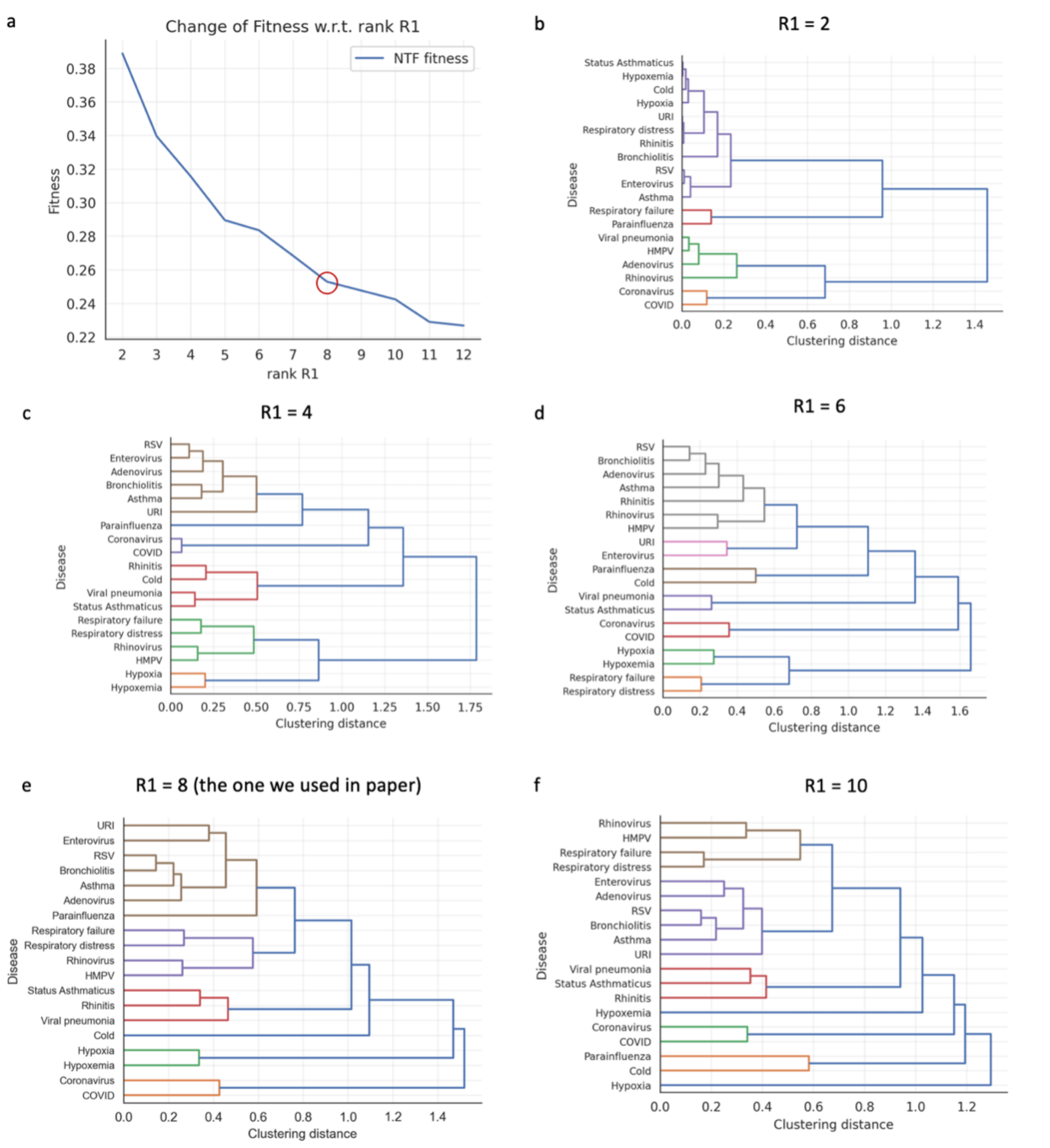


Figure 2 (supp). Ablation study on R_1. **a.** We plot the change of NTF fitness loss with respect to different choices of R_1. We can observe that the fitness loss is decreasing when R_1 becomes larger while the decreasing trend slows down gradually; **b.** The hierarchical clustering map with $R_1 = 2$; **c.** The hierarchical clustering map with $R_1 = 4$; **d.** The hierarchical clustering map with $R_1 = 6$; **e.** The hierarchical clustering map with $R_1 = 8$ (the one we used in the study); **f.** The hierarchical clustering map with $R_1 = 10$. We find that given a proper R_1(such as 8), the fitness loss can be reasonably low, and the output clustering map is also informative.

The goal of NTF decomposition is to capture the major low-rank information in the tensor $\boldsymbol{I}$ and treat the part that does not fit into the low-rank structure as noise and remove it. To model this, we ensure that each element $\boldsymbol{I}(i,j,k)$ is approximated by the Einstein summation of low-rank factors $\sum_{r=1}^{R_1} \boldsymbol{A}_1(i,r)\,\boldsymbol{A}_2(j,r)\boldsymbol{D}(k,r)$. Collectively, the objective is defined by summing over the squared residual of every element.

$$\boldsymbol{\mathcal{L}}_{residual} = \sum_{i=1}^{N}\sum_{j=1}^{T_1}\sum_{k=1}^{19}\left(\boldsymbol{I}(i,j,k) - \sum_{r=1}^{R_1}\boldsymbol{A}_1(i,r)\,\boldsymbol{A}_2(j,r)\boldsymbol{D}(k,r)\right)^2 \tag{1}$$

Additionally, we hope to extract the interpretable factors, which require (i) three factors containing non-negative elements (i.e., a common constraint for real count-based data); (ii) each disease representation being sparse (i.e., diseases have distinct temporal patterns). Thus, we add the non-negative constraints and the L_1 sparsity norm for disease representation $\boldsymbol{D}$.

$$\boldsymbol{\mathcal{L}}_{\geqslant 0}\colon [\boldsymbol{A}_1 \geqslant 0, \boldsymbol{A}_2 \geqslant 0, \boldsymbol{D} \geqslant 0] \tag{2}$$

$$\boldsymbol{\mathcal{L}}_{sparsity} = ||\boldsymbol{D}||_1 \tag{3}$$

To form the final objective, we add ridge regularization [42] to control the scale of the factors for preventing numerical errors.

$$\boldsymbol{\mathcal{L}}_{ridge} = ||\boldsymbol{A}_1||_F^2 + ||\boldsymbol{A}_2||_F^2 + ||\boldsymbol{D}||_F^2 \tag{4}$$

which is the sum of L_2 Frobenius norm of the factor matrices.

The final objective is a weighted sum of the above objectives in consideration of the non-negative constraint $\boldsymbol{\mathcal{L}}_{\geqslant 0}$,

$$\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{L}}_{residual} + \lambda_1 \cdot \boldsymbol{\mathcal{L}}_{sparsity} + \lambda_2 \cdot \boldsymbol{\mathcal{L}}_{ridge}$$

Here, λ_1, λ_2 are two hyperparameters. In this study, λ_1 is used in the proximal step (show below), and we find that a large $\lambda_1$will significantly hurt the decomposition fitness (when $\lambda_1 > 5 \times 10^{-1}$) while a small λ_1 cannot guarantee a sparse result (when $\lambda_1 < 1 \times 10^{-2}$). Thus, we choose $\lambda_1 = 1 \times 10^{-1}$ in our study. For λ_2, prior CP tensor decomposition research [40,41] usually set it between 1×10^{-8} to 1×10^{-5}. We did not find obvious performance change when increasing λ_2 from 1×10^{-8} to 1×10^{-5}. Thus, we set $\lambda_2 = 1 \times 10^{-5}$ throughout the experiments.

NTF optimization procedure

To optimize the overall objective function as well as the non-negative constraint, we use the proximal [43, 44] alternating least squares (ALS) algorithm [45] to update the factors $\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{D}$ sequentially. The algorithm runs for several iterations to converge, and each iteration consists of three sub-iterations. In the beginning of the algorithm, we initialize three factors $\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{D}$ all by independent identically uniform $\sim [0,1]$ distributed values. We describe one iteration below (including three sub-iterations, and they run in a sequence):

Sub-iteration 1: we fix the value of $\boldsymbol{A}_2, \boldsymbol{D}$ and only update $\boldsymbol{A}_1$. In this case, $\boldsymbol{A}_1$ are the only parameters and the current objective function becomes:

$$\boldsymbol{\mathcal{L}}_1 = \sum_{i=1}^{N}\sum_{j=1}^{T_1}\sum_{k=1}^{19}\left(\boldsymbol{I}(i,j,k) - \sum_{r=1}^{R_1}\boldsymbol{A}_1(i,r)\,\boldsymbol{A}_2(j,r)\boldsymbol{D}(k,r)\right)^2 + \lambda_2 \cdot ||\boldsymbol{A}_1||_F^2 \tag{5}$$

This objective is a quadratic form with respect to $\boldsymbol{A}_1$, and we can use the closed-form solution to obtain a new $\boldsymbol{A}_1$ by

$$\boldsymbol{A}_1 \leftarrow \boldsymbol{I}^{(1)}\,(\boldsymbol{A}_2 \odot \boldsymbol{D})(\boldsymbol{A}_2'\boldsymbol{A}_2 * \boldsymbol{D}'\boldsymbol{D} + \lambda_2)^{-1} \tag{6}$$

Here, $\boldsymbol{I}^{(1)}$ is the matricized tensor along the first dimension, $\odot$ is the matrix Khatri-Rao product, * is the matrix Hadamard product, $'$ is the matrix transpose, and $(\cdot)^{-1}$ is the matrix inverse operation. All the operations are well explained in this work [41]. In the end, we clip the element value of $\boldsymbol{A}_1$ into the range $[10^{-5}, +\infty)$ to meet the non-negative constraint and prevent the zero-division issue.

Sub-iteration 2: we fix the value of $\boldsymbol{A}_1, \boldsymbol{D}$ and only update $\boldsymbol{A}_2$. Like sub-iteration 1, currently, $\boldsymbol{A}_2$ are the only parameters and the current objective function is again a quadratic form.

$$\boldsymbol{\mathcal{L}}_2 = \sum_{i=1}^{N}\sum_{j=1}^{T_1}\sum_{k=1}^{19}\left(\boldsymbol{I}(i,j,k) - \sum_{r=1}^{R_1}\boldsymbol{A}_1(i,r)\,\boldsymbol{A}_2(j,r)\boldsymbol{D}(k,r)\right)^2 + \lambda_2 \cdot ||\boldsymbol{A}_2||_F^2 \qquad (7)$$

We can use the closed-form solution to obtain a new $\boldsymbol{A}_2$ by

$$\boldsymbol{A}_2 \leftarrow \boldsymbol{I}^{(2)}(\boldsymbol{A}_1 \odot \boldsymbol{D})(\boldsymbol{A}_1'\boldsymbol{A}_1 * \boldsymbol{D}'\boldsymbol{D} + \lambda_2)^{-1} \qquad (8)$$

Where $\boldsymbol{I}^{(2)}$ is the matricized tensor along the second dimension. Again, after obtaining the new $\boldsymbol{A}_2$, we apply the $[10^{-5}, +\infty)$ clipping.

Sub-iteration 3: we fix the value of $\boldsymbol{A}_1, \boldsymbol{A}_2$ and only update $\boldsymbol{D}$. This procedure is a bit more complicated since the current objective is not a quadratic form.

$$\boldsymbol{\mathcal{L}}_3 = \sum_{i=1}^{N}\sum_{j=1}^{T_1}\sum_{k=1}^{19}\left(\boldsymbol{I}(i,j,k) - \sum_{r=1}^{R_1}\boldsymbol{A}_1(i,r)\,\boldsymbol{A}_2(j,r)\boldsymbol{D}(k,r)\right)^2 + \lambda_1 \cdot ||\boldsymbol{D}||_1 + \lambda_2 \cdot ||\boldsymbol{D}||_F^2 \qquad (9)$$

To obtain the new $\boldsymbol{D}$, we first use the pesudo closed-form solution,

$$\boldsymbol{D} \leftarrow \boldsymbol{I}^{(3)}(\boldsymbol{A}_1 \odot \boldsymbol{A}_2)(\boldsymbol{A}_1'\boldsymbol{A}_1 * \boldsymbol{A}_2{}'\boldsymbol{A}_2 + \lambda_2)^{-1} \qquad (10)$$

Where $\boldsymbol{I}^{(3)}$ is the matricized tensor along the third dimension. We later apply the proximal step to enforce the sparsity on each disease representation separately (each row of $\boldsymbol{D}$**)**

$$\boldsymbol{d}_i \leftarrow \boldsymbol{d}_i - \lambda_1 \cdot max(\boldsymbol{d}_i) \cdot \boldsymbol{1} \qquad (11)$$

Here, $max(\boldsymbol{d}_i)$ means the max value in vector $\boldsymbol{d}_i$, and $\boldsymbol{1}$ is a constant vector with all 1 as the entry and has the same shape as $\boldsymbol{d}_i$. Again, we clip the new $\boldsymbol{D}$ within the range $[10^{-5}, +\infty)$.

Disease hierarchical clustering

We can see in **Fig. 2b (main)** that the disease representation is indeed sparse, and each disease is only connected to one or a few patterns. For obtaining the disease hierarchical clustering, we first normalize the representation $\boldsymbol{d}_i$ by the sum of the scores from all ranks (now the vector $\boldsymbol{d}_i$ sums up to 1) and then apply the agglomerative clustering approach [29] to get the clustering distance. We hope that diseases with similar patterns can be clustered closely. Therefore, we further apply the Ward variance minimization algorithm [39] on the distances and group the disease one at a time to form the clustering hierarchy in **Fig. 2c (main)**.

Analysis 2: Non-negative Tensor Factorization for Location Clustering

In this task, we want to analyze and understand the regional disparity of RSV trends. We apply the same NTF methods on the claim data $\boldsymbol{I}$ (only the RSV), the CDC data $\boldsymbol{C}$, and the google trend data $\boldsymbol{G}$ separately, while the data are formatted as matrices (two-dimensional tensors) here, one dimension is for locations (county or state), and another dimension is for the timeline (in week granular). The decomposition results will reflect different RSV progression patterns as well as the corresponding location clusters. We use medical claims data as example below.

NTF objective on claim data

Formally, given the RSV portion from disease tensor $\boldsymbol{I}(:,:,1) \in \mathbb{R}^{N\times T_1}$ (assume the index of RSV is 1 in the third disease dimension) up to 08/20/2022, the rank-R_2 NTF approach will decompose it into: a county representation matrix $\boldsymbol{L} \in \mathbb{R}^{N\times R_2}$, and a week representation matrix

$\boldsymbol{B}_1 \in \mathbb{R}^{T_1 \times R_2}$, following the CPD model. Here, each row of $\boldsymbol{L}$ is the location (i.e., county) represenation $\boldsymbol{l}_j \in \mathbb{R}^{R_2}$, and each column of $\boldsymbol{B}_1$ is one distinct RSV progresion pattern. In the experiment, we use $R_2 = 3$ as the number of clusters. A large R_2 would give finer granular clustering results with more location clusters, and we leave it for future work.

As the objective, we ensure that each element $\boldsymbol{I}(i, j, 1)$ is approximated by the Einstein summation of low-rank factors $\sum_{r=1}^{R_2} \boldsymbol{L}(i, r)\, \boldsymbol{B}_1(j, r)$. Collectively, the objective is defined by summing over the squared residual of every element.

$$\boldsymbol{\mathcal{L}}_{residual} = \sum_{i=1}^{N} \sum_{j=1}^{T_1} \left(\boldsymbol{I}(i, j, 1) - \sum_{r=1}^{R_2} \boldsymbol{L}(i, r)\, \boldsymbol{B}_1(j, r) \right)^2 \tag{12}$$

Similar to Analysis 1, we hope the final factors can be interpretable, which requires (i) $\boldsymbol{L}, \boldsymbol{B}_1$ containing non-negative elements; (ii) each location representation being sparse (i.e., locations have distinct temporal patterns). Thus, we add the non-negative constraints and the L_1 sparsity norm. Additionally, we also use the ridge regularization to control the factor scale.

$$\boldsymbol{\mathcal{L}}_{\succcurlyeq 0}: [\boldsymbol{L} \succcurlyeq 0, \boldsymbol{B}_1 \succcurlyeq 0] \tag{13}$$

$$\boldsymbol{\mathcal{L}}_{sparsity} = ||\boldsymbol{L}||_1 \tag{14}$$

$$\boldsymbol{\mathcal{L}}_{ridge} = ||\boldsymbol{L}||_F^2 + ||\boldsymbol{B}_1||_F^2 \tag{15}$$

The final objective is a weighted sum of the above objectives in consideration of the non-negative constraint $\boldsymbol{\mathcal{L}}_{\succcurlyeq 0}$,

$$\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{L}}_{residual} + \alpha_1 \cdot \boldsymbol{\mathcal{L}}_{sparsity} + \alpha_2 \cdot \boldsymbol{\mathcal{L}}_{ridge} \tag{16}$$

Here, α_1, α_2 are two hyperparameters. Similar to the setting of Analysis 1, we set $\alpha_1 = 1 \times 10^{-1}$, $\alpha_2 = 1 \times 10^{-5}$ in the experiments.

NTF optimization procedure

The optimization procedures are similar to Analysis 1 as well. In the beginning of the algorithm, we initialize three factors $\boldsymbol{L}, \boldsymbol{B}_1$ all by independent identically uniform $\sim [0,1]$ distributed values. One iteration is described below (including two sub-iterations, run in a sequence):

Sub-iteration 1: we fix the value of $\boldsymbol{L}$ and only update $\boldsymbol{B}_1$. Then, $\boldsymbol{B}_1$ are the only parameters and the current objective function becomes:

$$\boldsymbol{\mathcal{L}}_1 = \sum_{i=1}^{N} \sum_{j=1}^{T_1} \left(\boldsymbol{I}(i, j, 1) - \sum_{r=1}^{R_2} \boldsymbol{L}(i, r)\, \boldsymbol{B}_1(j, r) \right)^2 + \alpha_2 \cdot ||\boldsymbol{B}_1||_F^2 \tag{17}$$

This objective is a quadratic form with respect to $\boldsymbol{B}_1$, and we can use the closed-form solution to obtain a new $\boldsymbol{B}_1$ by

$$\boldsymbol{B}_1 \leftarrow \boldsymbol{I}(:, :, 1)\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{L} + \alpha_2)^{-1} \tag{18}$$

In the end, we clip the element value of $\boldsymbol{B}_1$ into the range $[10^{-5}, +\infty)$ to meet the non-negative constraint and prevent the zero-division issue.

Sub-iteration 2: we fix the value of $\boldsymbol{B}_1$ and only update $\boldsymbol{L}$. The objective is not quadratic.

$$\boldsymbol{\mathcal{L}}_2 = \sum_{i=1}^{N} \sum_{j=1}^{T_1} \left(\boldsymbol{I}(i, j, 1) - \sum_{r=1}^{R_2} \boldsymbol{L}(i, r)\, \boldsymbol{B}_1(j, r) \right)^2 + \alpha_1 \cdot ||\boldsymbol{L}||_1 + \alpha_2 \cdot ||\boldsymbol{L}||_F^2 \tag{18}$$

To obtain the new $\boldsymbol{L}$, we first use the pseudo closed-form solution,

$$\boldsymbol{L} \leftarrow \boldsymbol{I}(:,:,1)'\boldsymbol{B}_1(\boldsymbol{B}_1'\boldsymbol{B}_1 + \alpha_2)^{-1} \tag{19}$$

Later, we apply the proximal step to enforce the sparsity on each location representation separately (each row of $\boldsymbol{L}$)

$$\boldsymbol{l}_j \leftarrow \boldsymbol{l}_j - \alpha_1 \cdot max(\boldsymbol{l}_j) \cdot \mathbf{1} \tag{20}$$

Here, $max(\boldsymbol{l}_j)$ means the max value in vector $\boldsymbol{l}_j$, and $\mathbf{1}$ is a constant vector with all 1 as the entry and has the same shape as $\boldsymbol{l}_j$. Again, we clip the new $\boldsymbol{L}$ within the range $[10^{-5}, +\infty)$.

Arguments of the maxima clustering

After the NTF optimization, we select the arguments of the maxima (argmax) index of each location representation $\boldsymbol{l}_i$ to get the location clustering. In total, three location clusters and three distinct progression patterns are identified, shown in **Fig. 3a (main)**. For the Google search data $\boldsymbol{G} \in \mathbb{R}^{51 \times T_3}$ and the CDC survillance data $\boldsymbol{C} \in \mathbb{R}^{12 \times T_2}$, we apply the same methods to obtain the clusters and the progression patterns, independently, provided in Supplemental materials (**Fig. 3 (supp)**, **Fig. 4 (supp)**). It is interesting that the location clusters and three RSV trend patterns are aligned across three different data sources.

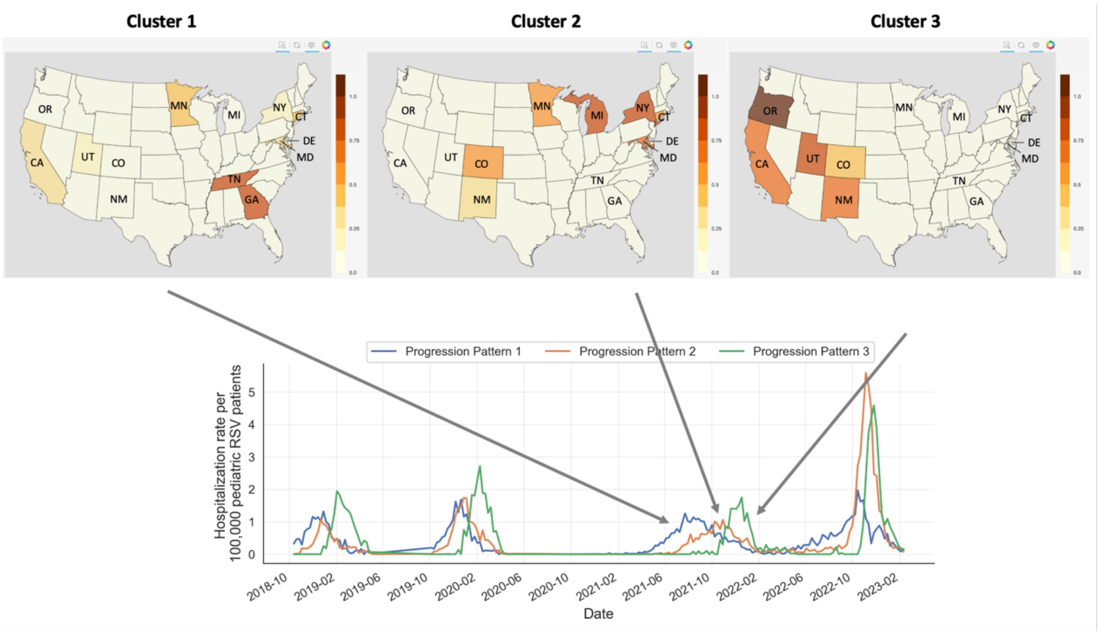


Figure 3 (supp). RSV progression location clusters (12 states from CDC surveillance data). The values in the upside map are the clustering intensity. The y-axis in the downside plot shows the hospitalization rate per 100,000 pediatric patients for RSV.

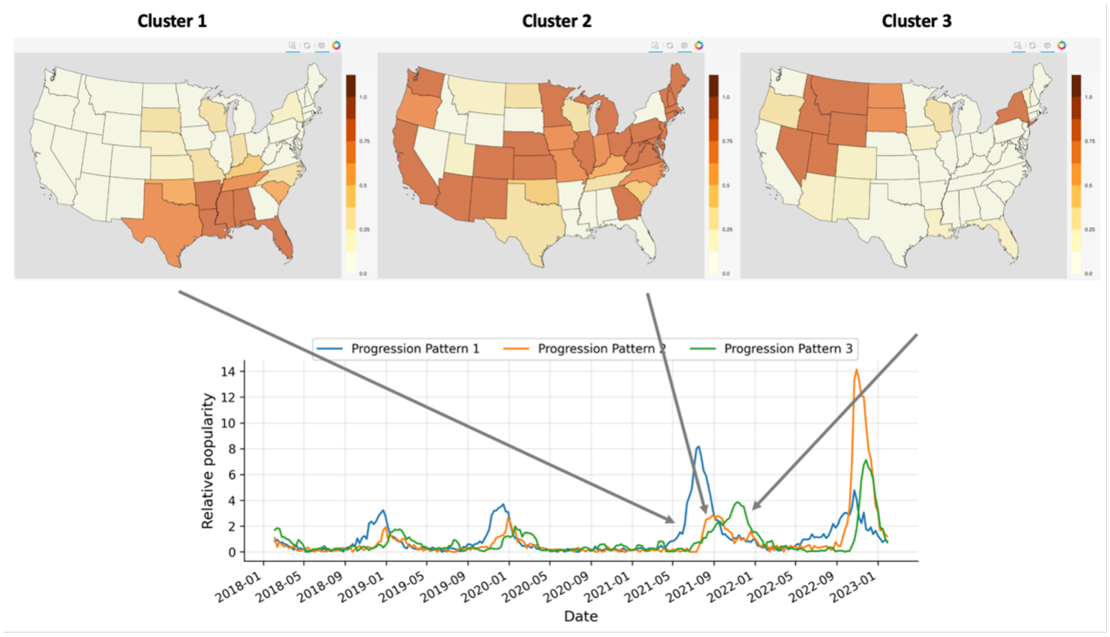


Figure 4 (supp). RSV progression location clusters (51 states in Google RSV search trends). The values in the upside map are the clustering intensity. The y-axis in the downside plot shows the relative popularity of RSV search.

The Lists of Static Features

We provide a detailed list of static feature matrices $\boldsymbol{M}_3 \in \mathbb{R}^{N\times 14}$ and $\boldsymbol{M}_4 \in \mathbb{R}^{N\times 39}$ in **Table 1 (supp)** and **Table 2 (supp)**. For the distance matrix $\boldsymbol{M}_1 \in \mathbb{R}^{N\times N}$ and the mobility distance matrix $\boldsymbol{M}_2 \in \mathbb{R}^{N\times N}$, their rows and columns both refer to the N counties.

Table 1 (supp). All dimensions (14) in demographics matrix $\boldsymbol{M}_3 \in \mathbb{R}^{N\times 14}$.

1	Overall population
2	0 – 17 age group population
3	18 – 64 age group population
4	65 plus age group population
5	Black-skin group population
6	White-skin group population
7	Asian population
8	Hispanic population
9	Non-Hispanic population
10	Number of Physicians
11	Number of Hospitals
12	Number of ICU Beds
13	Average amount of income per family
14	Unemployment rate

Table 2 (supp). All dimension (39) in COVID, hospitalization and vaccination matrix $\boldsymbol{M}_4 \in \mathbb{R}^{N\times 39}$.

1	Total vaccine shots	21	Total booster shots rate
2	Total 1st shots	22	Total Pfizer 1st shots rate

3	Total 2nd shots
4	Total booster shots
5	Total Pfizer 1st shots
6	Total Pfizer 2nd shots
7	Total Pfizer booster shots
8	Total Moderna 1st shots
9	Total Moderna 2nd shots
10	Total Moderna booster shots
11	Total Johnson 1st shots
12	Total Johnson booster shots
13	Total Pfizer TS 1st shots
14	Total Pfizer TS 2nd shots
15	Total Pfizer TS booster shots
16	Total Pfizer TS10 1st shots
17	Total Pfizer TS10 2nd shots
18	Total vaccine shots rate
19	Total 1st shots rate
20	Total 2nd shots rate
23	Total Pfizer 2nd shots rate
24	Total Pfizer booster shots rate
25	Total Moderna 1st shots rate
26	Total Moderna 2nd shots rate
27	Total Moderna booster shots rate
28	Total Johnson 1st shots rate
29	Total Johnson booster shots rate
30	Total Pfizer TS 1st shots rate
31	Total Pfizer TS 2nd shots rate
32	Total Pfizer TS booster shots rate
33	Total Pfizer TS10 1st shots rate
34	Total Pfizer TS10 2nd shots rate
35	Total COVID counts up to 05/01/2022
36	Total of in beds patients up to 05/01/2022
37	Total of in beds COVID patients up to 05/01/2022
38	Total of ICU patients up to 05/01/2022
39	Total of ICU COVID patients up to 05/01/2022

Feature Sets of Each Model

To provide more information on how the baseline model is implemented, we list the feature sets each model used in **Table 3 (supp)**.

Table 3 (supp). Feature sets of each baseline models. Basically, all models have the same amount of data for fairness (except ARIMA model which cannot use the static features). Our model outperforms other baselines due to that we explicitly extracted disease and location embeddings from the disease trends and input the prediction month as timing information.

ARIMA	RSV disease trends
LSTM	County level static features RSV and other disease trends
Transformer	County level static features RSV and other disease trends
XGBoost	County level static features RSV and other disease trends
STAN	County level static features RSV and other disease trends
HOIST	County level static features RSV and other disease trends
Our TAP-RSV	County level static features RSV and other disease trends Timing of prediction (our new features) Disease tensor representation (our new features, can be extracted from disease trends in (2)) County tensor representation (our new features, can be extracted from disease trends in (2))

Additional Experimental Results

Additional result 1: five-year trends of 19 pediatric diseases

We aggregate the disease data over all counties and plot their overall trend on the timeline. Different diseases present distinct trends, while many diseases share a similar pattern, such as Bronchiolitis, Cold, Hypoxemia, RSV. The results are shown in **Figure 1 (supp).**

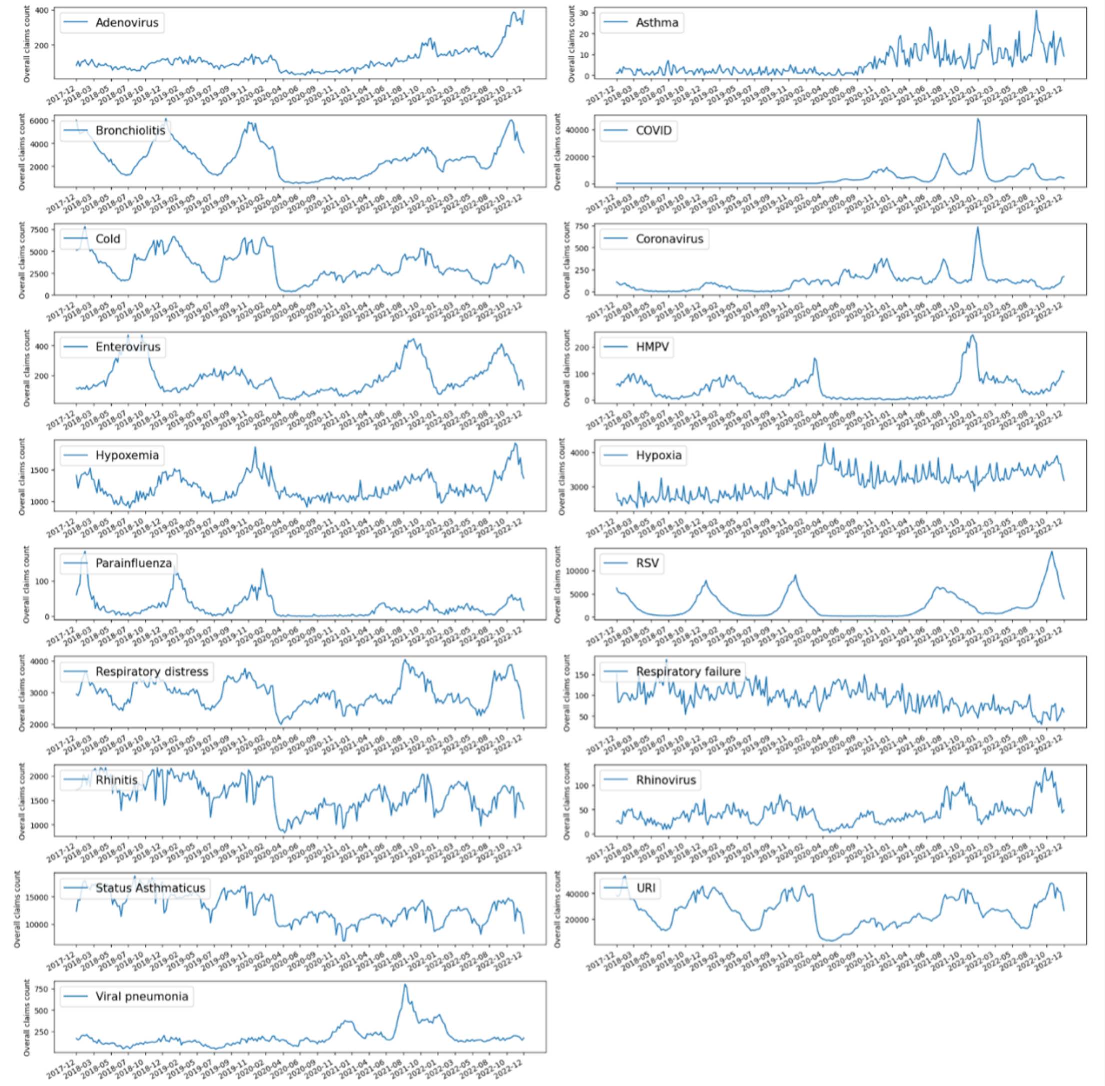


Figure 1 (supp). The progression trends of all 19 diseases. We aggregate the disease count over all counties and plot their progression trends on the timeline. Different diseases present distinct trends, while many diseases share a similar pattern, such as Bronchiolitis, Cold, Hypoxemia, RSV.

Additional result 2: varying the rank R_1 in disease hierarchical clustering

The NTF approach can factorize the claim tensor $\boldsymbol{I}$ into three low-dimensional matrices, and the rank choice of 8 is justified below. We conduct ablation studies on R_1 for Analysis 1. We plot the change of NTF fitness loss in **Figure 2 (supp)(a)**. The fitness loss is decreasing when R_1 becomes larger while the decreasing trend slows down gradually. We also show the hierarchical clustering results given $R_1 = 2, 4, 6, 8, 10$ in **Figure 2 (supp)(b)-(f)**, while $R_1 = 8$ is used in our experiments (**Figure 2 (supp)(e)**). We find that the clustering results of $R_1 = 2, 4$ are of low quality (for example, Bronchiolitis and RSV are not the nearest) compared to $R_1 = 6, 8, 10$. One possible reason is that 19 diseases can present more than 4 general patterns (thus, $R_1 \leq 4$ can fail), as shown in **Figure 1 (supp)**. While $R_1 = 6, 8, 10$ seem to capture all typical trends, and their clustering maps make clinical sense to some extent without major differences. Note that since we are clustering for 19 diseases, R_1 should be kept smaller than 19. We choose $R_1 = 8$ in the experiments because it has a relatively low fitness loss, and the output clustering are informative.

Additional result 3: location clusters and patterns for Google and CDC data

Location clusters and patterns for Google and CDC data are shown in **Figure 2 (supp), Figure 3 (supp)**.

Additional result 4: better performance on longer prediction window

To evaluate the performance capacity of baselines and our model, we fix the observation window up to 08/20/2022 and change the prediction window size to 4 weeks, 8 weeks, 12 weeks, and 16 weeks. We want to show that our model is still advantageous in longer window prediction (up to the entire 4-month season prediction at once). In this experiment, we do not include the ARIMA model, as it trains very slowly (using the *pmdarima.auto_arima* implementation) and is not comparable to other baseline models. We illustrate the prediction setting illustrations in **Figure 4 (supp).**

For this new setting, we use the same four metrics and plot the comparison in **Figure 5 (supp).** Each plot represents one metric, and each curve (associated with the error bar) represents one model. Based on the results, we can tell that the performance of all models degrades with longer prediction window, and our model consistently outperforms the baselines in all scenarios. LSTM appears to be a very strong and stable baseline. Note that we find Transformer model gives abnormal behaviors once the prediction window becomes 08/20 -> 10/15. Although it seems to have decent MSE and MAE metric, we can analyze from the extremely low PCC values (< 0.1) that in these scenarios, the Transformer model actually outputs random results (which may likely to be all near 0).

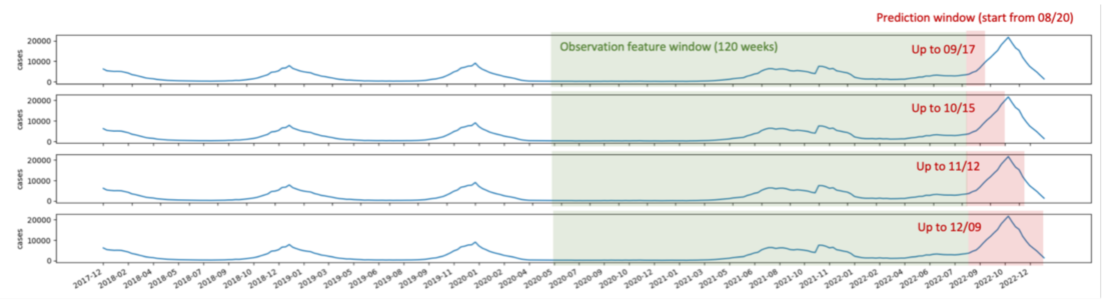


Figure 5 (supp). Illustration of longer window prediction for 2022-2023 season (4, 8, 12, 16 weeks ahead)

Additional result 5: distribution of true and predicted county-level case counts

To support our statement in **Figure 4 (main)(b)** that "the relative scale of the prediction is generally smaller than the true scale". We plot the x-log distribution of the true and predicted county-level case counts in **Figure 6 (supp).** As we can observe that the predicted distributions are more towards the left compared to the true case count distribution, which verifies our statement.

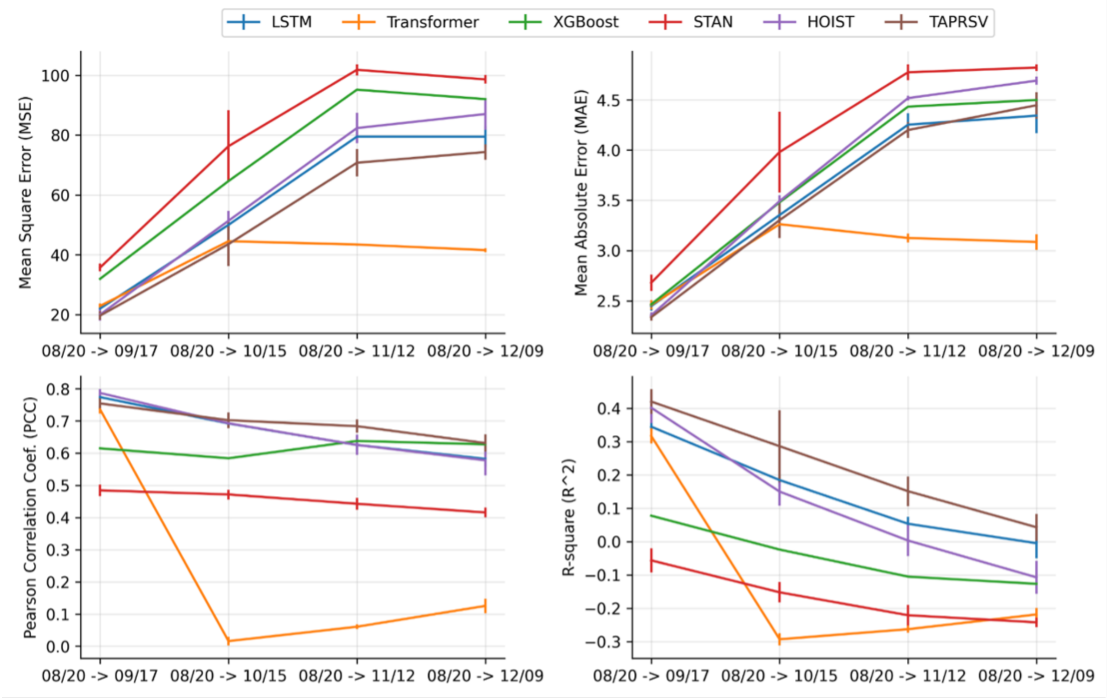


Figure 6 (supp). Performance comparison in longer window prediction. The performance of all models degrades with longer prediction window, and our model consistently outperforms the baselines in all scenarios. The Transformer model fails starting at the second prediction. Its PCC values are close to 0, which means that the output becomes random.

Additional result 6: ablation study on model feature components

Given the performance, we analyze the contributions of different feature components (i.e., disease embedding, location embedding, month embedding, static features, and the observation time-series) in our TAP-RSV model. We consider the following model variants:

- **TAPRSV-No-Disease-Emb:** our model without the disease embedding $\boldsymbol{d}_i$ $(i = 1, \dots, 19)$;
- **TAPRSV-No-Static-Feature:** our model without the static features $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \boldsymbol{M}_4$;
- **TAPRSV-No-Location-Emb:** our model without location embedding $\boldsymbol{l}_j$ $(j = 1, \dots, 2334)$;
- **TAPRSV-No-Timing:** our model without the month embedding $\boldsymbol{M}(m)$;
- **TAPRSV-No-Time-series:** our model without the disease time-series.

For each model variants, we retrain them from scratch under five initializations with different random seeds. We compare our full TAPRSV model with the variants on the prediction window 10/15/2022 – 11/12/2022 since this window includes the peak of RSV. The performance is shown as bar charts in **Figure 7 (supp)**. The first model is our full TAPRSV, and we roughly sort the model variants by their feature importance order (same for all figures). Obviously, the time-series sequence is the most important feature by default. The model will degrade a lot without time-series on all metrics. We also find that the timing (i.e., the month representation) and the location embeddings are two other important factors in the prediction, which can be explained from the location clustering perspective. Refer to the clustering map in **Figure 2 (main)(a), Figure 2 (supp), Figure 3 (supp)**, we can tell that the future RSV progression pattern of a county or a state might be uniquely identified by the location cluster and the timing.

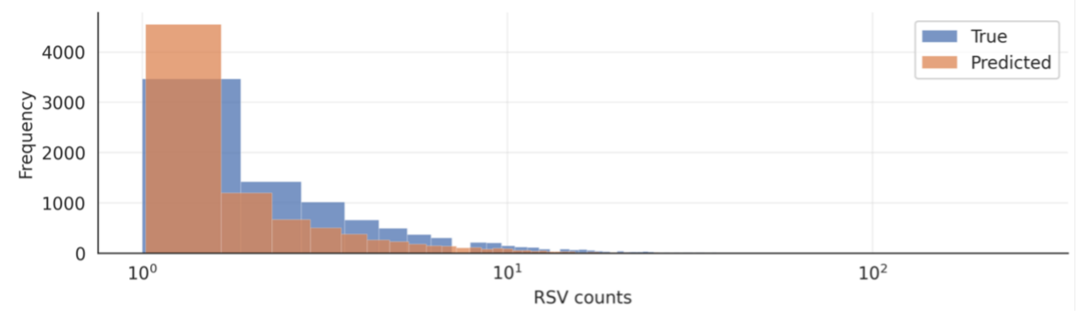


Figure 7 (supp) | x-log distribution of true and predicted county-level case counts | We can observe that the predicted distributions are more towards the left compared to the true case count distribution.

Additional Results 7: p-values of one-sided T-test for Table 1 (main)

The p-values of one-sided T-test for **Table 1 (main)** is shown in **Table 4 (supp)**.

Table 4 (supp). P-values of one-sided T-test of Table 1 (main). Format: 4 decimal precisions in scientific notation. The results are generated under 5 random seeds. We can observe that most of the p-values are smaller than 0.05, meaning that the performance gain of our TAP-RSV model is generally significant over the baselines. Among all baselines, the HOIST model shows the best performance. There are some cases that our TAP-RSV model shows better performance over the baselines, however, the p-values are larger than 0.05. We mark them by shade. "/" means that HOIST model is better than TAP-RSV in this metric.

Window	Model	MSE	MAE	PCC	R^2
08/20 ➔ 09/17	ARIMA	7.1887E-09	1.0164E-09	1.2598E-14	5.0936E-08
	LSTM	4.7233E-04	1.6080E-03	4.5601E-02	2.5594E-04
	Transformer	1.6685E-04	1.2140E-03	1.3864E-04	7.6581E-05
	XGBoost	4.3296E-09	3.9258E-04	9.8868E-11	3.0978E-10
	STAN	7.9293E-09	1.6716E-05	1.8372E-10	1.9347E-09
	HOIST	2.2492E-01	4.4111E-02	/	3.2351E-01
09/17 ➔ 10/15	ARIMA	2.7831E-01	1.1002E-05	9.1656E-15	1.5390E-09
	LSTM	3.2731E-05	2.6077E-05	1.4142E-04	2.2312E-05
	Transformer	1.4479E-05	1.1229E-05	5.6260E-05	6.6992E-06
	XGBoost	1.2640E-08	1.1622E-09	3.1737E-09	4.4004E-09
	STAN	1.7523E-09	1.6079E-10	1.3792E-10	4.3610E-10
	HOIST	2.3959E-02	2.3559E-04	6.0403E-02	1.0702E-02
10/15 ➔ 11/12	ARIMA	1.6544E-03	1.0217E-08	9.5779E-11	5.6173E-09
	LSTM	1.1905E-05	3.6428E-03	4.9555E-06	9.3181E-06
	Transformer	5.1119E-06	2.0060E-05	1.3783E-05	5.7073E-06
	XGBoost	5.6750E-10	6.0760E-12	1.7774E-04	1.5647E-10
	STAN	1.5008E-10	7.7191E-12	5.4203E-08	3.8955E-11
	HOIST	9.3976E-03	3.2501E-03	1.3736E-01	5.0287E-03
11/12 ➔ 12/09	ARIMA	9.0695E-10	1.4510E-05	1.8718E-11	3.2344E-15
	LSTM	2.0936E-05	1.2723E-03	3.9307E-05	3.3909E-03
	Transformer	8.7420E-06	4.9299E-03	2.5154E-05	3.5115E-04
	XGBoost	4.3591E-11	3.1213E-08	7.8752E-06	1.9077E-09
	STAN	4.5037E-10	3.7976E-07	7.4526E-08	2.6607E-09

	HOIST	7.2957E-02	4.6526E-02	1.4069E-01	1.9835E-01