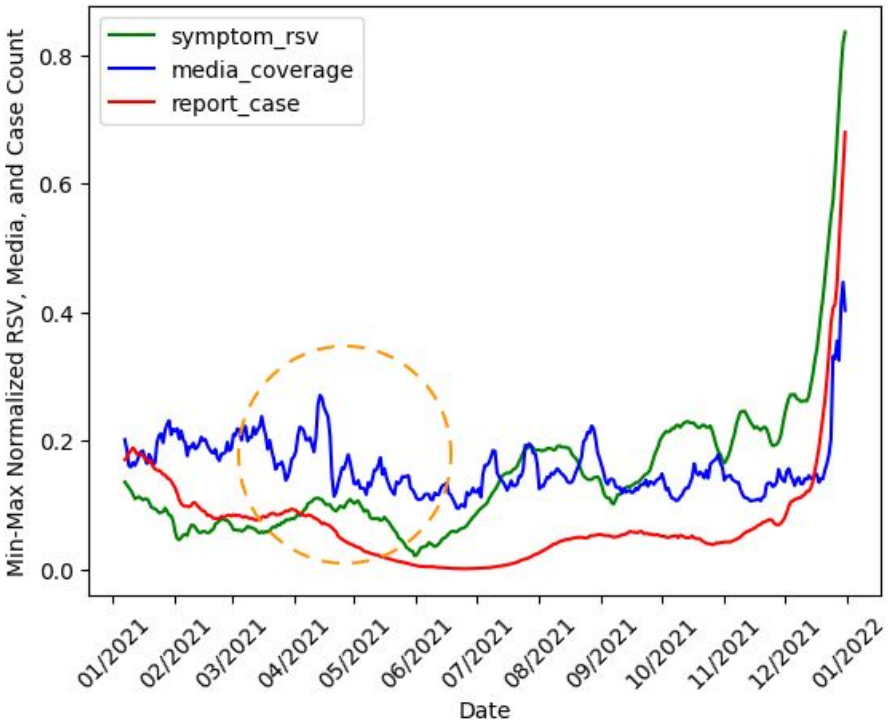
Removing Media Coverage Bias from COVID-19 Symptom Search Trends Data

Siddhartha Vemuri, Shen En Chen, Mohit Aggarwal

Table of Contents

Motivation



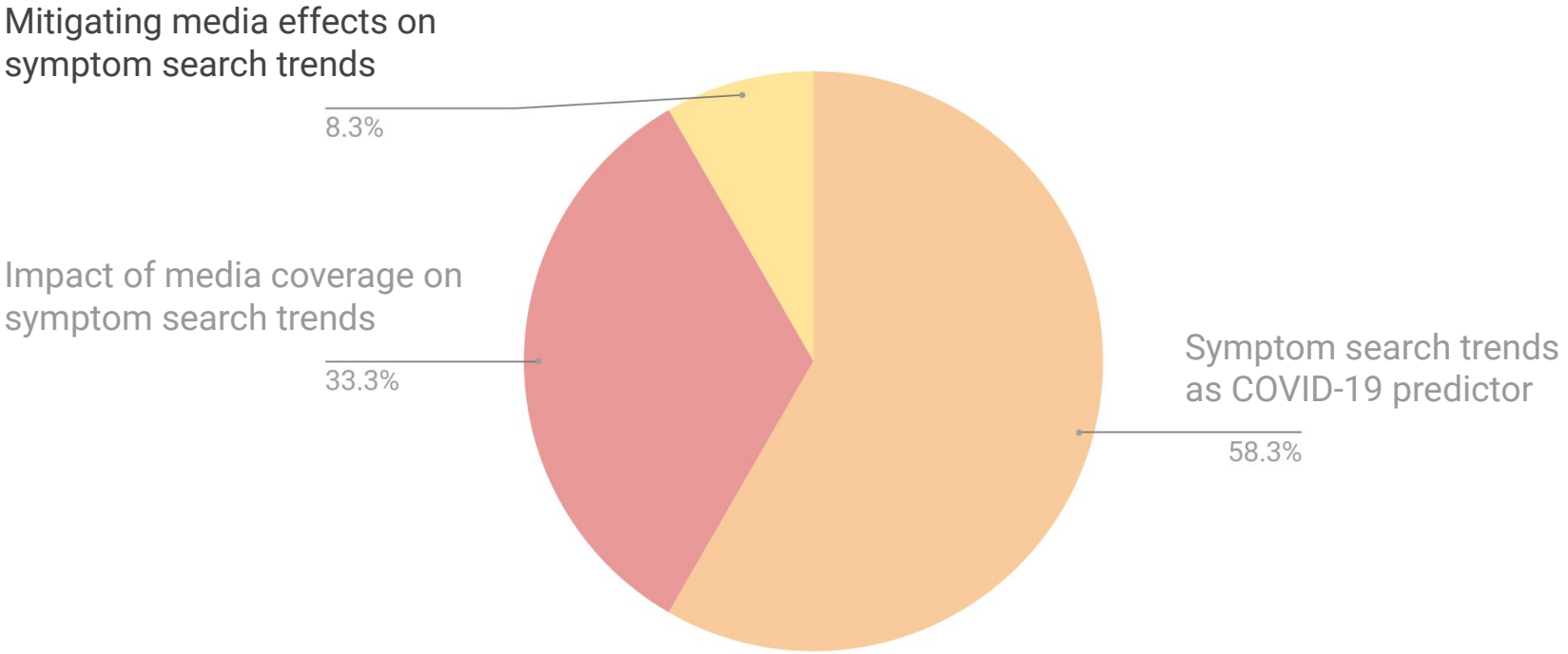
Symptom search trend

Media coverage

Reported COVID Cases

Spikes in symptom trend are sometimes caused by other confounding factors, such as media coverage.

Related Work



Problem Statement

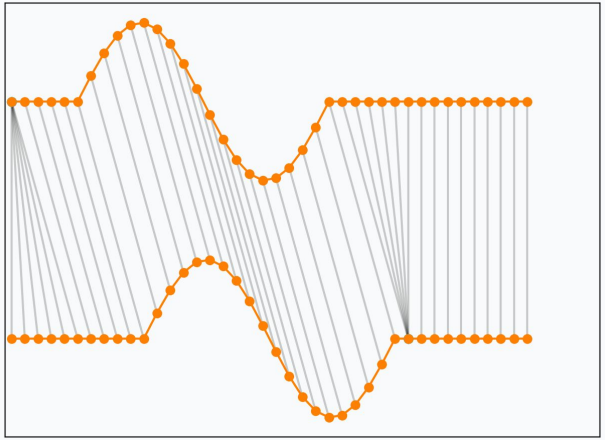
Remove the impact of media coverage

on the Google relative search volumes of COVID-19 symptoms

with various approaches.

Evaluation Metrics

- Root-mean-square error (RMSE)
  → curve similarity
- Pearson correlation
  → general trends
- Spearman rank correlation
  → general trends



- Minimum distance for dynamic time warping (DTW)
  → similarity temporal alignment

Data Collection

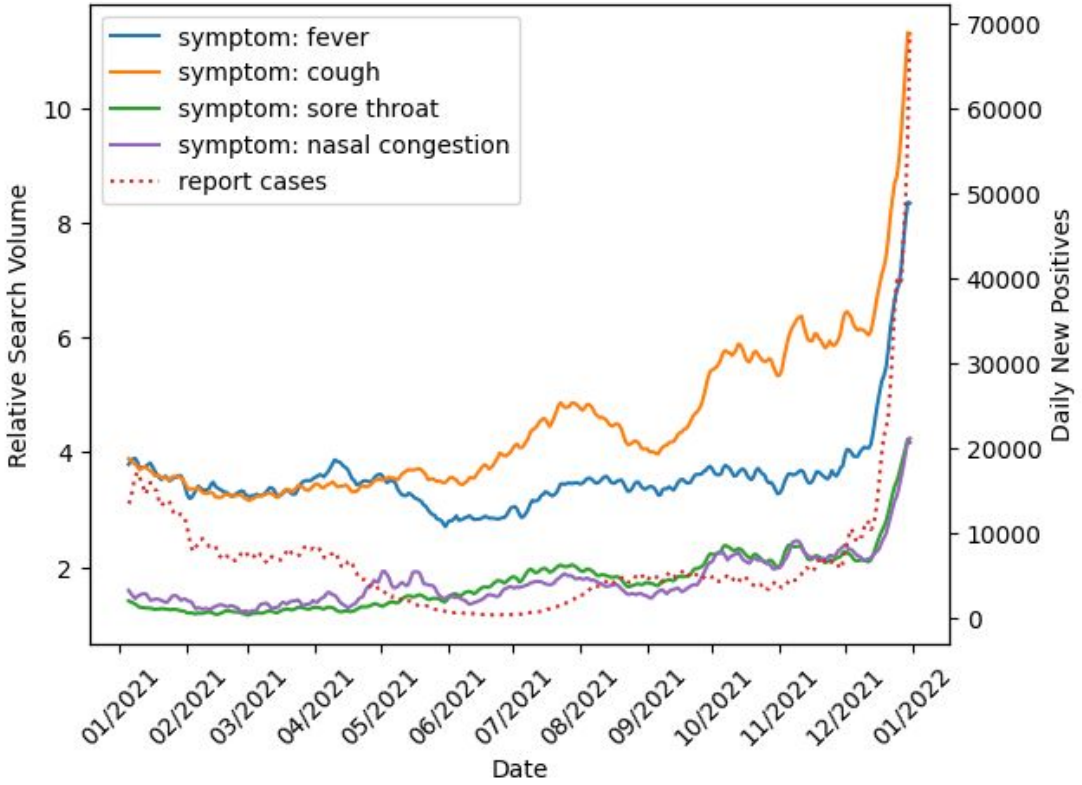
- COVID-19 case counts from NY State Statewide COVID-19 Testing API [1]
- RSV for several relevant symptoms from Google Search Trends [2]

  ```
  ['Fever', 'Chills', 'Cough', 'Shortness of breath', 'Sore throat',
  'Headache', 'Fatigue', 'Muscle weakness', 'Anosmia', 'Ageusia', 'Nasal
  congestion', 'Nausea', 'Vomiting', 'Diarrhea']
  ```

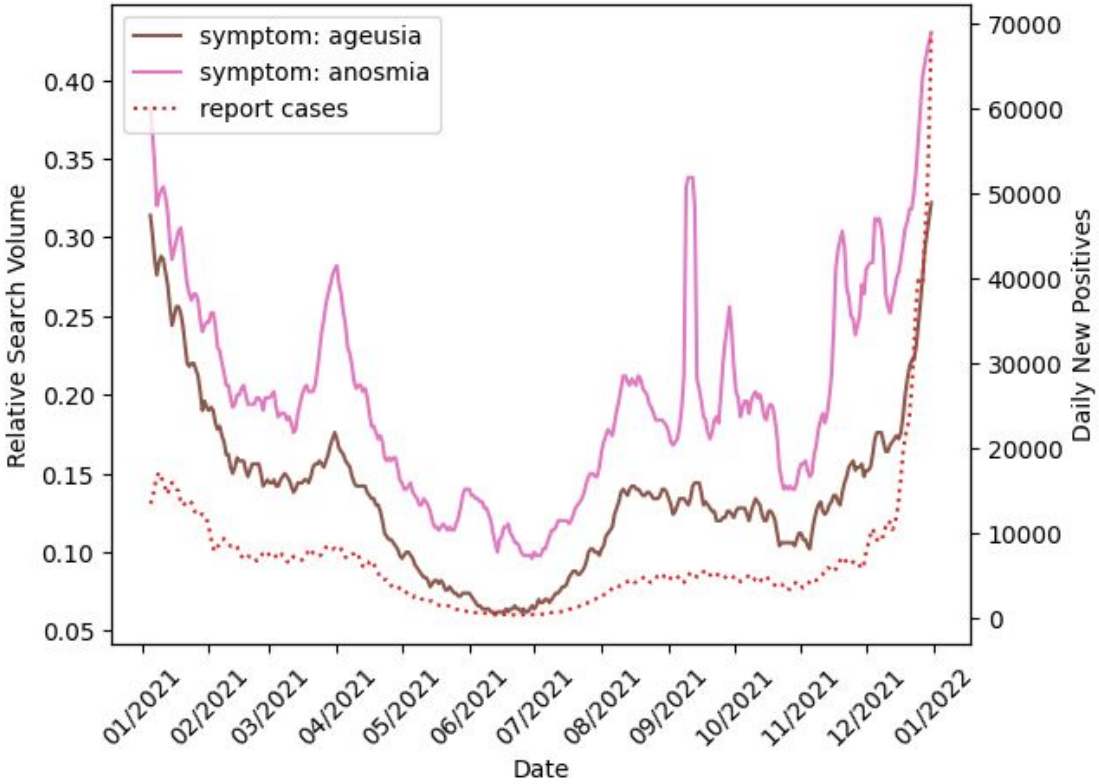
- Media coverage for each of the above symptoms from the MediaCloud Database [3]

[1] New York State Department of Health. 2022. New York State Statewide COVID-19 Testing.
[2] Google LLC. 2020. Google COVID-19 Search Trends symptoms dataset.
[3] Media Cloud. 2022. Media Cloud is an open-source platform for media analysis.

Preliminary Analysis



Preliminary Analysis



Weighted Signal Aggregation

- Create combined RSV and combined media coverage signals by
  - Applying symptom frequency values as weights to individual symptom data
  - Summing all values to create one aggregated signal
- Proceed with experiments using weighted aggregated RSV and media coverage signals

Ageusia	Anosmia	Chills
0.20	0.20	0.63
Cough	Diarrhea	Fatigue
0.84	0.38	0.62
Fever	Headache	Muscle Weakness
0.80	0.59	0.63
Nasal Congestion	Nausea	Shortness of Breath
0.10	0.34	0.57
Sore Throat	Vomiting	
0.40	0.13	

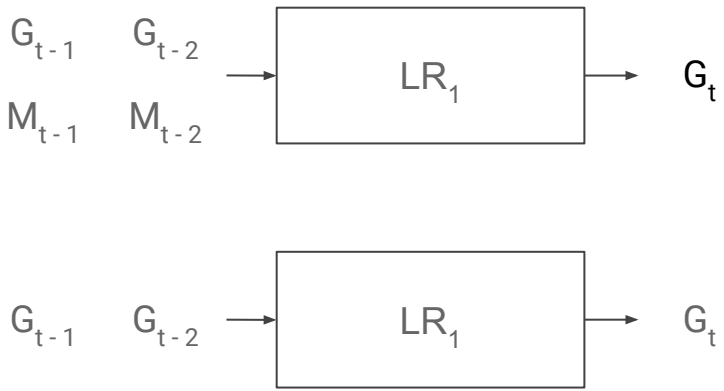
Table 2: Reported frequencies of symptoms in study of COVID patients

Burke et al. 2020. Symptom Profiles of a Convenience Sample of Patients with COVID-19 — United States, January–April 2020. Morbidity and Mortality Weekly Report (2020).

Baseline: Lampos et al.

Vasileios Lampos, Maimuna S. Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei Hamada, Maolebogeng X. Rangaka, Rachel A. McKendry, and Ingemar J. Cox. 2021. Tracking COVID-19 using online search. npj Digital Medicine 4, 11 (Feb 2021), 1–11.

Step 1:



G_t = symptom search trend at time t

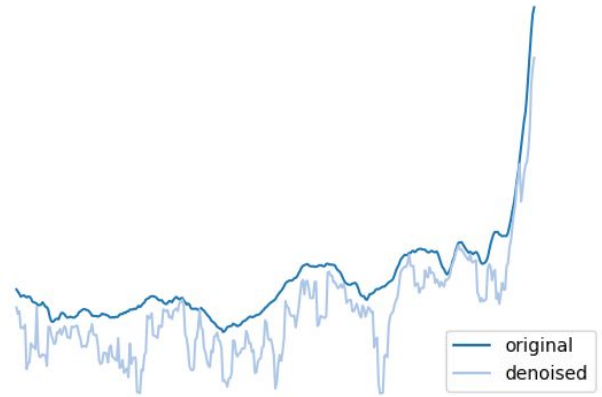
M_t = media coverage at time t

LR = linear regression model

Step 2:

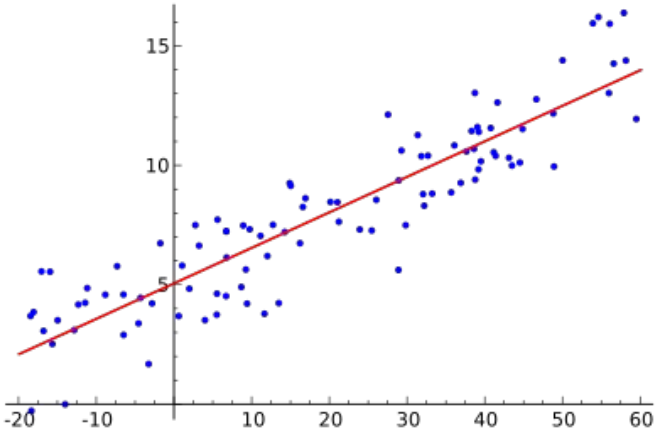
```
for t in max_time:
      if LR1 has smaller error:
            G_t = G_t * err(LR_1) / err(LR_2)
      else:
            G_t = G_t
```

Step 3:



Our Approach: Linear Regression

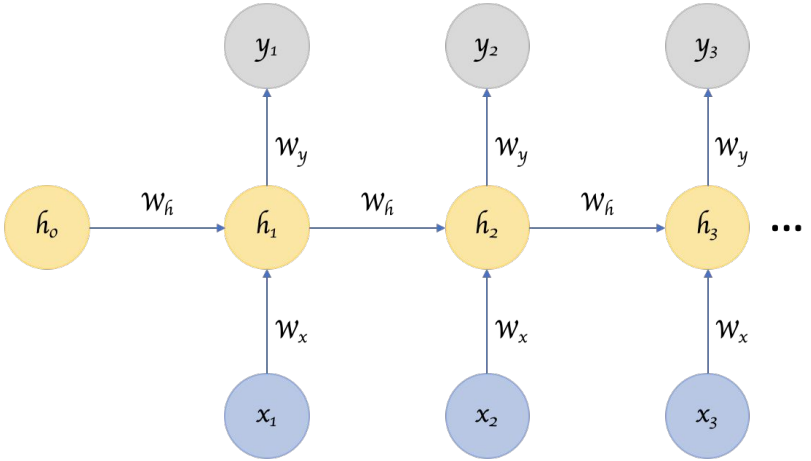
- Evaluate the following versions of a linear regression model over lead times from k=[1, 30] days
- Models
  - **Given** RSV data for day `i`,
    predict case counts for day `i + k`
  - **Given** RSV and case count data for day `i`,
    predict case counts for day `i + k`
  - **Given** RSV and media data foro day `i`,
    predict case counts for day `i + k`
  - **Given** RSV, media data, and case counts for day `i`,
    predict case counts for day `i + k`



- Intuition
  - Allow the model to find an ideal weight to apply to media coverage data to remove its influence from RSV data used as a predictor for case counts

Our Approach: Recurrent Neural Networks

- Intuition
  - Recurrent neural networks can learn more sequential dependencies between all three input variables when forecasting the case counts.
- Input
  - Symptom RSV (past)
  - Media coverage (past)
  - Case counts (past)
- Output
  - Case counts (current)

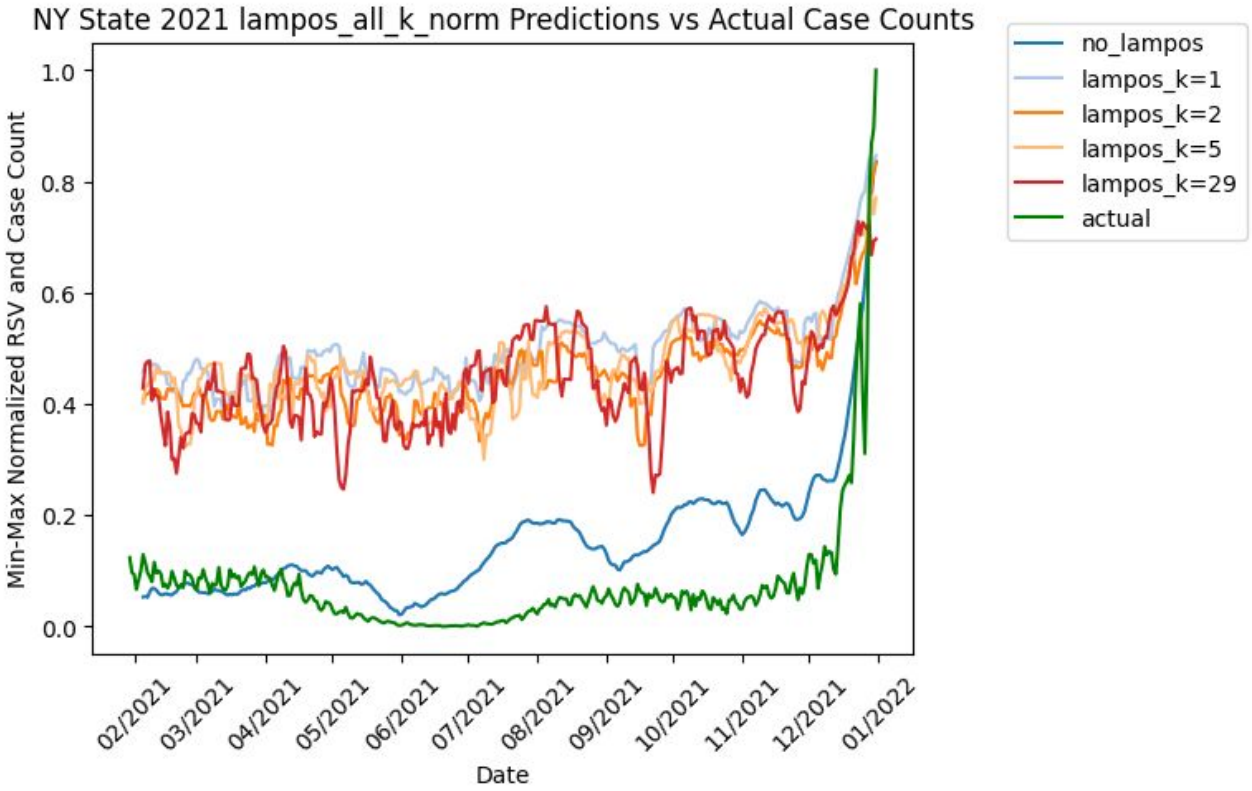


Results: Lampos et al.

Lag	RMSE	Pearson	Spearman	DTW
0	11894.05	0.75 ($2.58 \cdot 10^{-69}$)	0.19 ($1.20 \cdot 10^{-4}$)	10.29
1	11884.87	**0.52** ($2.24 \cdot 10^{-26}$)	0.18 ($3.42 \cdot 10^{-4}$)	125.73
2	11886.48	0.49 ($4.12 \cdot 10^{-13}$)	0.17 ($1.43 \cdot 10^{-3}$)	111.36
5	11870.53	0.43 ($2.14 \cdot 10^{-17}$)	**0.19** ($2.42 \cdot 10^{-4}$)	116.94
29	**11668.07**	0.29 ($2.96 \cdot 10^{-8}$)	0.14 ($8.44 \cdot 10^{-3}$)	**84.57**

Table 3: Performance of Lampos et al. [10] with different lags. Lag = 0 days represents the performance without applying denoising. For Pearson and Spearman rank correlations, we format the results as *score (p-value)*.

Results: Lampos et al.



Results: Linear Regression

Lead Time	With Media Data	With Current Case Count	RMSE	Pearson	Spearman	DTW
1	Yes	Yes	2359.95	0.98 ($1.73 \cdot 10^{-119}$)	0.91 ($2.02 \cdot 10^{-69}$)	2.31
1	No	Yes	2674.86	0.97 ($2.75 \cdot 10^{-108}$)	0.92 ($7.53 \cdot 10^{-74}$)	2.32
1	Yes	No	3153.83	0.98 ($3.29 \cdot 10^{-119}$)	0.86 ($3.40 \cdot 10^{-53}$)	4.07
1	No	No	3353.88	0.97 ($5.95 \cdot 10^{-110}$)	0.87 ($9.89 \cdot 10^{-58}$)	4.16
7	Yes	Yes	3575.71	0.95 ($9.53 \cdot 10^{-92}$)	0.86 ($3.81 \cdot 10^{-53}$)	3.20
7	No	Yes	3554.38	0.95 ($6.58 \cdot 10^{-92}$)	0.87 ($9.69 \cdot 10^{-56}$)	3.62
7	Yes	No	4675.03	0.94 ($4.49 \cdot 10^{-84}$)	0.81 ($1.68 \cdot 10^{-41}$)	4.35
7	No	No	4684.23	0.94 ($4.23 \cdot 10^{-83}$)	0.84 ($6.39 \cdot 10^{-47}$)	4.25

Table 4: Performance of linear regression for lead times of 1 and 7, including and excluding media data, and including and excluding current day case count data (all combinations tested). When looking at the Pearson and Spearman columns, note that values are formatted as *score (p-value)*.

Results: Recurrent Neural Networks

Sequence Length	With Media Data	RMSE	Pearson	Spearman	DTW
2	Yes	5185.83	0.96 ($3.09 \cdot 10^{-43}$)	0.94 ($2.07 \cdot 10^{-34}$)	0.90
2	No	6322.54	0.95 ($3.47 \cdot 10^{-37}$)	0.93 ($2.77 \cdot 10^{-32}$)	0.97
4	Yes	5468.47	0.97 ($7.42 \cdot 10^{-44}$)	0.94 ($7.71 \cdot 10^{-34}$)	1.04
4	No	4937.88	0.97 ($1.29 \cdot 10^{-43}$)	0.94 ($1.03 \cdot 10^{-33}$)	0.88
5	Yes	6501.54	0.97 ($6.94 \cdot 10^{-44}$)	0.94 ($4.93 \cdot 10^{-35}$)	1.37
5	No	6143.85	0.97 ($1.09 \cdot 10^{-43}$)	0.93 ($7.23 \cdot 10^{-32}$)	1.18
7	Yes	5580.93	0.96 ($3.89 \cdot 10^{-39}$)	0.93 ($7.42 \cdot 10^{-30}$)	0.87
7	No	11861.85	0.95 ($4.15 \cdot 10^{-34}$)	0.90 ($1.05 \cdot 10^{-23}$)	1.85
10	Yes	7416.88	0.97 ($1.11 \cdot 10^{-39}$)	0.95 ($6.81 \cdot 10^{-33}$)	1.00
10	No	16315.65	0.87 ($1.12 \cdot 10^{-20}$)	0.79 ($5.55 \cdot 10^{-15}$)	2.75
14	Yes	7042.55	0.95 ($7.25 \cdot 10^{-32}$)	0.94 ($1.12 \cdot 10^{-28}$)	0.81
14	No	15608.15	0.77 ($4.85 \cdot 10^{-13}$)	0.83 ($2.38 \cdot 10^{-16}$)	2.03

Table 5: Performance of the recurrent neural network. We format the Pearson and Spearman columns as *score (p-value)*.

Discussion: Model Performances

- **Lampos et al.**
  - Does not yield much performance improvements on the metrics we evaluated.
- **Linear regression**
  - 1-day lag gives us the best results across most of the metrics evaluated
  - No significant increase in performance with media data
  - Not enough model complexity
  - Providing the case count data improves performance

Discussion: Model Performances

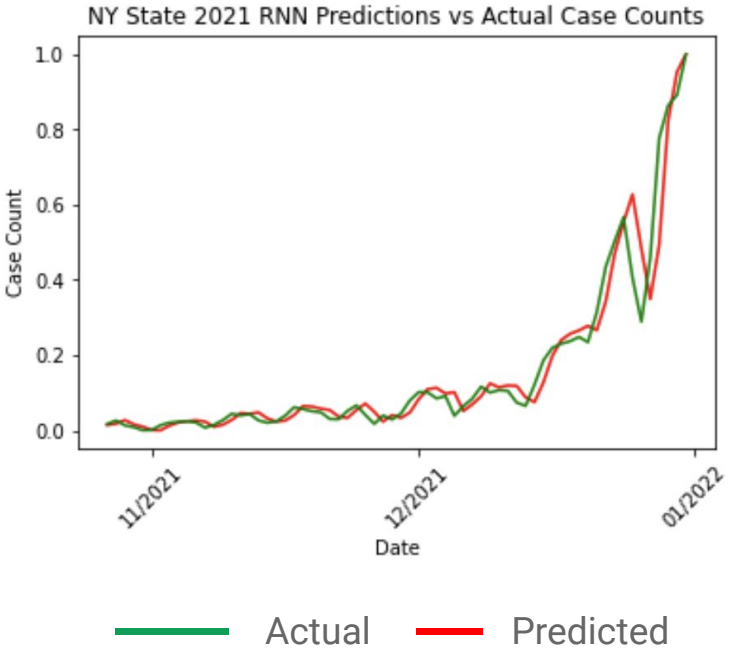
- **RNN**
  - Longer sequence might have increased noise in the input.
  - The Spearman rank correlation increases for the majority of the sequence lengths while the Pearson correlation remains relatively constant when media data is included.
  - The minimum distance for DTW using an RNN with media data included is much smaller than it would be for linear regression.
    - RNN has much better denoising capability than linear regression.

Discussion: Model Training

- Training RNN is fairly unstable, likely due to:
  - The low dimensionality of the features of the sequences
  - The amount of training data is not sufficient
- Using smaller datasets with more complicated designs was a bad idea. Large numbers of layers increased training time while also reducing training accuracy.
- The final product truly benefits from fine-tuning the hyper parameters, and we needed varied learning rates and units for various input sequences.

Discussion: Model Training

- The model could easily fail when the sequence length is too short due to insufficient data.
- A large number of epochs produces superior outcomes but is also prone to overfitting.



Conclusion

- **Lampos et al.** is both qualitative and quantitatively **insufficient** to achieve our objective of denoising media coverage from RSV
- **Linear Regression** works fairly well overall, but the single beta parameter assigned to the media coverage predictor variable is **insufficient** to remove its influence from RSV
- **RNN** performance is **highly variable based on architecture** and may fall into **undesirable local optima**, making it **difficult to evaluate** the strength of the approach

Overall: more work is needed to tune the RNN to properly evaluate its performance, but Lampos et al. and LR are conclusively insufficient

Future Work

- Apply our experiments to more common ILI data and other illnesses
  - Explore flu data to get more years of information
  - Explore rare illness outbreaks
- Continue exploring RNN architectures/tuning
  - Explore LSTMs or other methods that requires less data but with sufficient complexity
  - Investigate in separating symptom RSV as individual features
- Implement and study more metrics
  - Cross Correlation
  - Dynamical Conditional Correlation