“Nowcasting” Flu Hospitalizations using Google Search Data

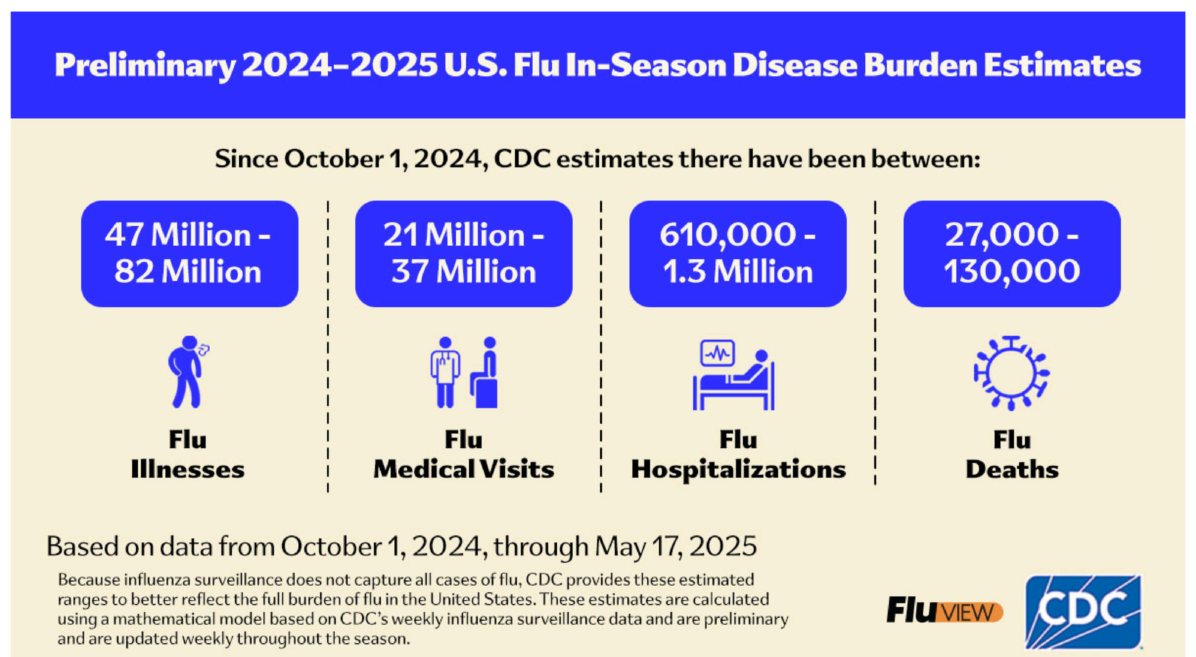
Kaito Hikino, Sunny Shi
Mentor: Shaoyang Ning

Introduction

The 2024–2025 influenza season was among the deadliest in recent history, with hospitalization rates matching or exceeding the highest totals in 15 years. Real-time tracking of flu activity helps public health officials make timely, life-saving decisions. Building on Professor Ning’s prior work using the ARGO (AutoRegression with Google data) framework for influenza-like illness (ILI), we extend this approach to predict weekly flu hospitalization rates.

Because hospitalization data from the National Healthcare Safety Network (NHSN) are only available after the COVID-19 pandemic, we tested strategies to overcome the limited training period. These include reducing required training length and imputing earlier hospitalization data.

Motivation



Google Search Data (Google Trend)

The CDC’s primary method of flu surveillance is its weekly influenza-like illness (ILI) reports, which provide estimated influenza-like illness prevalence and patient visits at the national, regional, and state levels. This system has one key drawback: **a 1 to 2-week reporting lag.** That delay happens because it takes time to collect, process, and aggregate clinical data. The ARGO model combines traditional data with real-time sources like Google search data to produce timely predictions.

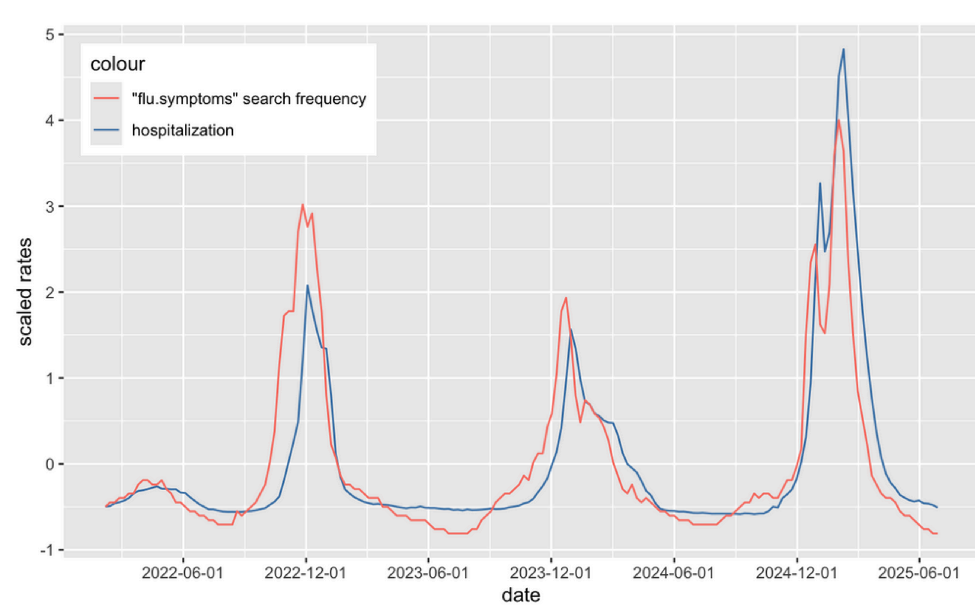


Figure 1. Google trend search volume v.s. flu hospitalization rate

Hospitalization Data Sources and Limitations

FluSurv-NET (prior to 2021-2022 flu season)

- limited resolution: reported to only one decimal place
- limited data: only during flu seasons and for certain states

NHSN (starting from 2021-2022 flu season)

- largest network of healthcare facilities for tracking healthcare-associated infections: includes approximately 25,000 healthcare facilities located throughout all 50 states, the District of Columbia, and Puerto Rico.
- **Due to this broader network, we also utilize NHSN data in our research.**

AutoRegression with GOogle search data

Let H_t' be the transformed hospitalization rate out of 100,000 at time t, formally $H_t' = logit((H_t + 0.1)/100)$. Then, the basic ARGO model is given by

$$H_t' = \mu_{H'} + \sum_{j=1}^{N} a_j H_{t-j}' + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

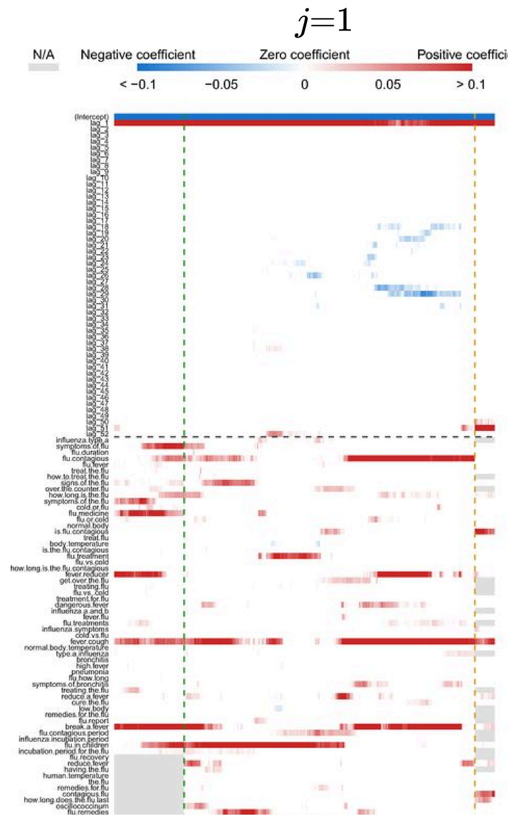


Figure 2. Coefficients using Lasso

We use **Lasso regression (L1 penalty)** to select predictors. On average, 14 Google search terms are selected for each prediction.

Since we only have three years of data, the chosen tuning parameters for ARGO do not work well with the hospitalization data. We used cross-validation to select the best parameters for hospitalization data.

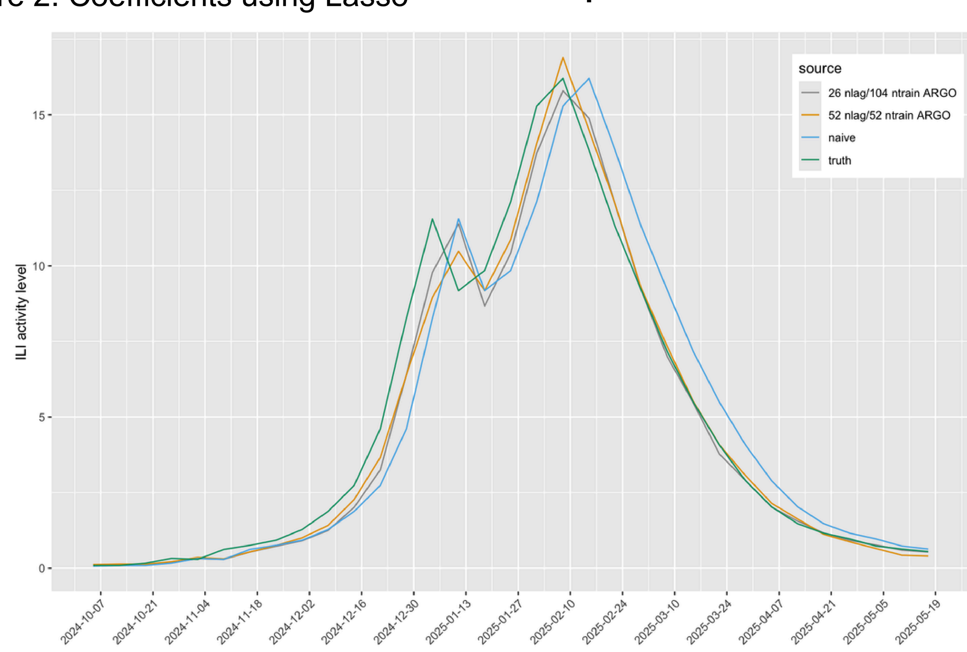


Figure 3. Predicted hospitalization rate

Parameter combinations and prediction relative MSE
Evaluation period: 2024/09/21 - 2025/05/24

Transformation	Lag Time	Training Time	Relative MSE
logit	26 weeks	78 weeks	0.2257
logit	32 weeks	78 weeks	0.2321
logit	52 weeks	52 weeks	0.2487

Figure 4. Best combinations of parameters

ARGO2 (Regional) and ARGOX (State)

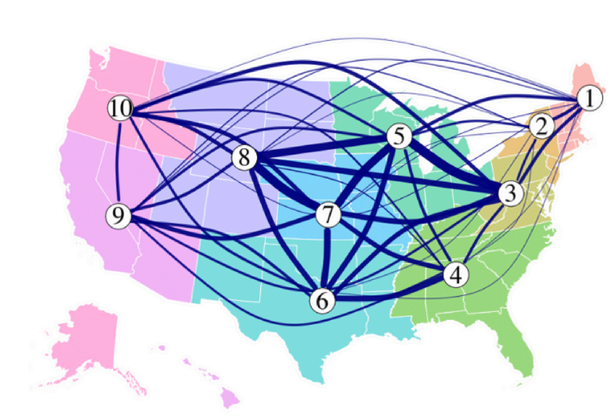


Figure 5. Spatial structure of regional data

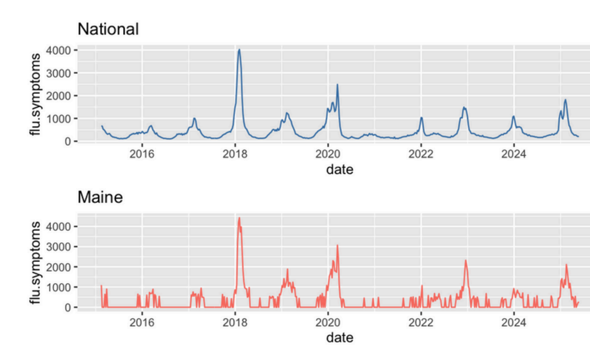


Figure 6. Sparsity of state-level data

ARGO2 and ARGOX operate on the regional and state level, respectively. Two changes are added to the original framework. As state and regional data **display strong spatial structure** due to movement between regions, our localized models use cross-regional/cross-state boosting. Additionally, since **Google search data is sparser** at these levels due to inadequate collection, ARGO2 & ARGOX utilize a weighted average of national and state data.

Imputations

Imputation was motivated by the limited three-year data and will be especially helpful in state-level predictions. We looked for linearity between ILI percentage and flu hospitalization rate under different transformations, including no transformation, logit[(x+0.1)/100], log(x+1), best normalize package selection, and ordered quantile normalization.

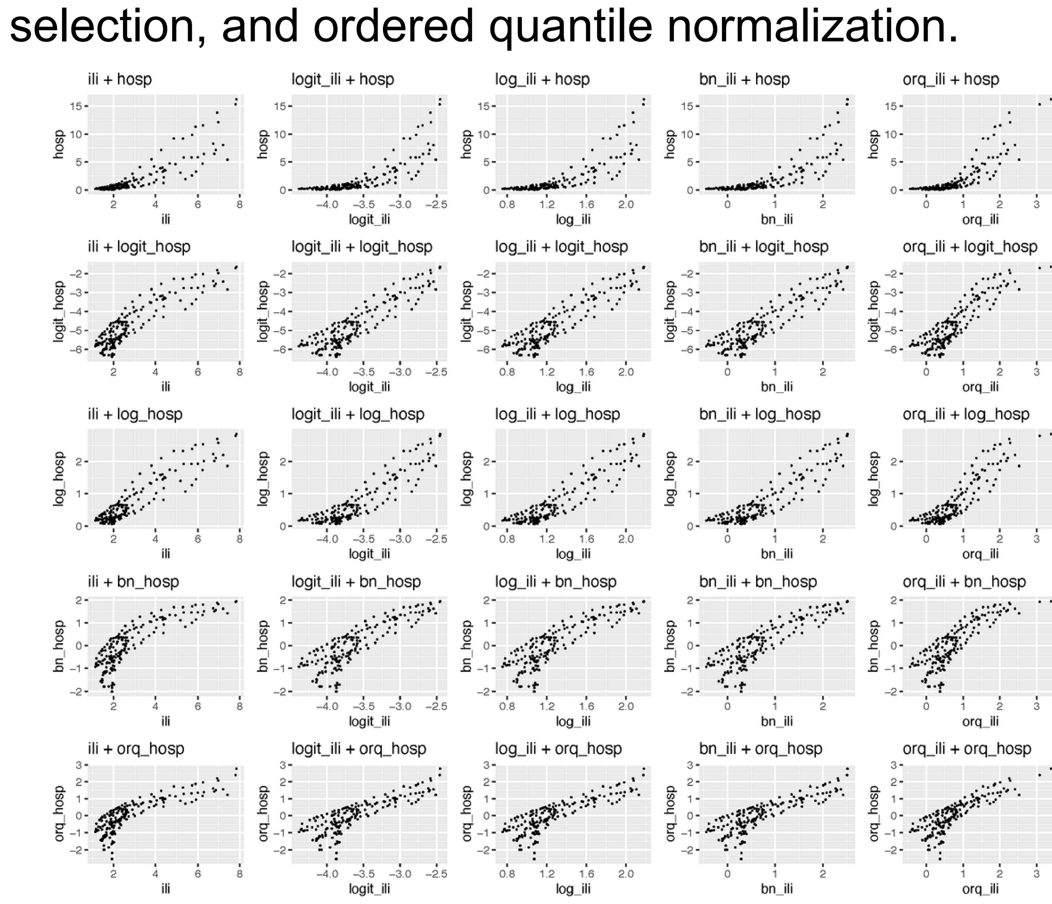


Figure 7. Scatterplots of transformed ILI against transformed hospitalization rate

However, **the ARGO model appears robust to different transformations.**

Imputation CV MSE and prediction relative MSE
Evaluation period: 2024/02/10 - 2025/05/24

ILI transformation	Hosp transformation	CV MSE	Prediction relative MSE
no imputation	no imputation	N/A	0.2396
original	ordered quantile	1.9649	0.3485
logit	logit	1.6423	0.3491
best normalize	best normalize	0.4256	0.3507
ordered quantil	ordered quantile	2.1804	0.3565
ordered quantile	original	1.1864	0.3787

Incorporating ILI

We used the past N weeks of logit-transformed ILI percentages as additional predictors in our model, formally

$$H_t' = \mu_{H'} + \sum_{j=1}^{N} \alpha_j H_{t-j}' + \sum_{i=1}^{K} \beta_i X_{i,t} + \sum_{k=1}^{N} \gamma_k ILI_{t-k}' + \epsilon_t, \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Relative MSE for full predicted period

ILI \ Hosp	lag 0	lag 26	lag 52
52 N-train			
lag 0	1.0866	0.2732	0.2378
lag 26	0.6117	0.2739	0.2377
lag 52	0.5814	0.2837	0.2378
lag 104	0.3234	0.2558	0.2432
104 N-train			
lag 0	0.8117	0.2763	0.1729
lag 26	1.1364	0.3201	0.1651
lag 52	1.0817	0.3242	0.1473
lag 104	1.0551	0.3489	0.1299

We need at least 104 weeks of ILI data to improve predictive accuracy, which would be an excessively large number of predictors.

Conclusion

We developed an ARGO framework for national hospitalizations with a consistent relative mean square error around 0.2. **Our ongoing research seeks to extend this framework to the state and regional level and evaluate our current framework on a longer timeframe.**

At the national level, even though imputation does not improve prediction accuracy, it may still provide valuable insight when comparing different models. We would like to know if any imputed data preserves the ranking of parameter combinations.

For state-level predictions, to address the lack of direct regional data, we plan to test two models: one with regional estimates as weighted averages of state-level data, and one that relies only on state-level predictors without regional boosting.

References and Acknowledgements

Meyer, A. G., Lu, Fred, Clemente, L., & Santillana, M. (2025). A prospective real-time transfer learning approach to estimate influenza hospitalizations with limited data. *Epidemics*, *50,* 100816.

Ning, S., Yang, S., Kou, S. C. (2019). Accurate regional influenza epidemics tracking using Internet search data. *Scientific Reports*, 9(1), 1-8.

Peterson RA (2021). “Finding Optimal Normalizing Transformations via bestNormalize.” *The R Journal*, 13(1), 310–329.

Yang, S., Ning, S., & Kou, S.C. (2021). Use Internet search data to accurately track state level influenza epidemics. *Scientific Reports*, 11, 4023(2021).

Yang, S., Santillana, M., & Kou, S.C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc. Natl. Acad. Sci. U.S.A.,* 112(47), 14473-14478.

We thank Professor Ning for guiding us through this research and Swarthmore College for providing us with the funding for this research.