Interplay and Predictive Power of COVID-19 Incidence Indicators in Virginia

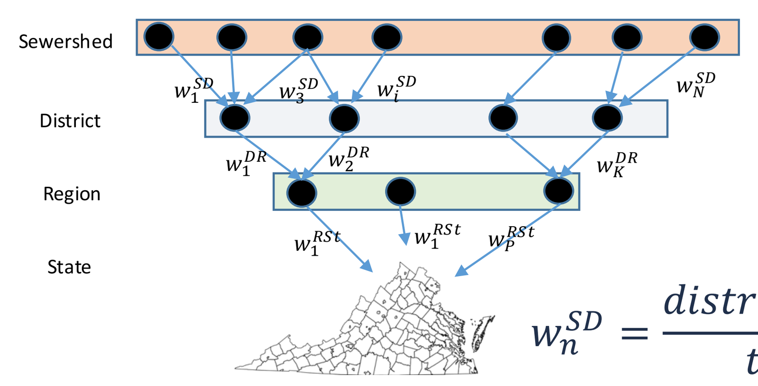
Gursharn Kaur, Aniruddha Adiga, Benjamin Hurt, Brian Klahn, Chen Chen, Andrew Warren, Srini Venkatramanan, Bryan Lewis, Madhav Marathe

Biocomplexity Institute, University of Virginia

Motivation

- **To investigate the relationship between wastewater data and various COVID-19 related indicators** such as Hospitalizations, Cases, COVID-Like Illness (CLI) by utilizing quantile (level-based) categorization.
- **Aim: to develop a predictive model for cases and hospitalizations (categories)**, while evaluating forecast performance based on different signals used in the model, time frames, and evaluation metrics.

Dataset Details & Spatiotemporal Alignment



Viral load aggregation from Sewersheds to districts

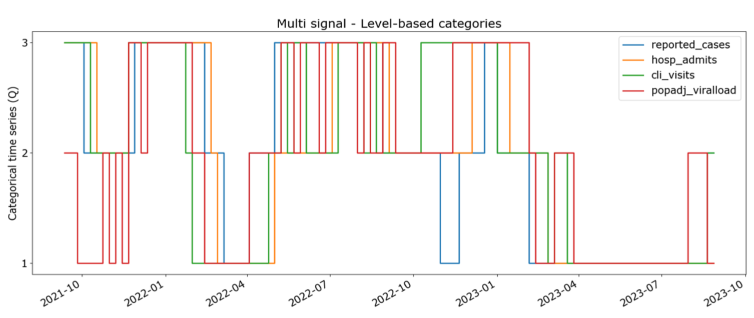
$$V_i(D,t) = \sum_{n=1}^{N} w_n^{SD} V_n^{S}(t)$$

$$w_n^{SD} = \frac{district\ pop.\ served\ by\ sewershed}{total\ district\ pop.\ served}$$

Categorical Indicators

Level-based categories: determining bin boundaries based on 3 quantiles.

- $1 = Low$,
- $2 = Med$,
- $3 = High$.



Test of association

- **Chi-square test**: to measure the significance of the association. H_0: Cases are independent of Viral loads
- Test Statistics:

$$\chi^2 = \sum_{i,j} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}};\ E_{i,j} = \frac{r_i c_j}{N}$$

Based on the observed data we get $\chi^2_{df=4} = 566.03$, p-value < 0.05, Cohen's $\omega = 0.52$

- **Chi-square test shows that cases and wastewater have a strong association.**

OBSERVED

Cases	Viral Load: Low	Med	High	Total
Low	**446**	192	68	706
Med	189	**281**	222	692
High	71	219	**398**	688
Total	706	692	688	2086

EXPECTED

Cases	Viral Load: Low	Med	High	Total
Low	239	234	233	706
Med	234	230	228	692
High	233	228	227	688
Total	706	692	688	2086

Multinomial regression formulation

We consider the multinomial regression model to forecast case and hospitalization categories. For instance, to fit a model

$$\eta(Q_{cases}(t)) = \beta_0 + \beta_1 Q_{VL}(t-1),$$

we train **2 logistic regression** models for $k = 2, 3$

$$\eta(Q_{cases}(t)) = \log\left(\frac{P(Q_{cases}(t) = k)}{P(Q_{cases}(t) = 1)}\right)$$
$$= \beta^k_{0,0} + \beta^k_{1,1} I_{QVL(t-1)1} + \beta^k_{1,2} I_{QVL(t-1)2} + \beta^k_{1,3} I_{QVL(t-1)3}$$

here $Q_{cases}(t)$ denotes the quantile-category for cases at time t and $I_{Q_{cases}(t-1)i}$ is the indicator variable for the event $\{Q_{cases}(t-1) = i\}$.

Forecast model: incorporating multiple signals

Experiment Setup:

- Time period: Sept 2021 to June 2023, starting with 20 weeks for model training, for each district.
- Real-time forecasting: updated quantile categories with each new data point.
- **Forecast horizon: 1-week ahead forecasts for all districts.**

The precision of combined district-level category forecasts is *significantly improved* by incorporating hospitalizations and CLI into the model.

- ✓ 11% improvement in average precision
- ✓ 13% increase in accuracy within the "High" category

Model 1: $\eta(Q_{cases}(t)) = \beta_0 + \beta_1 Q_{VL}(t-1)$

Model 1		Predicted Categories: Low	Med	High	Total
Actual categories	Low	**446** 63.2%	192 27.7%	68 9.9%	706 33.8%
	Med	189 26.8%	**281** 40.6%	222 32.3%	692 33.2%
	High	71 10%	219 31.7%	**398** 57.8%	688 33%
		706 100%	692 100%	688 100%	2086

Average precision = 53.93%

Model 2: $\eta(Q_{cases}(t)) = \beta_0 + \beta_1 Q_{VL}(t-1) + \boldsymbol{\beta_2 Q_{hosp}(t-1)} + \boldsymbol{\beta_3 Q_{CLI}(t-1)}$

Model 2		Predicted Categories: Low	Med	High	Total
Actual categories	Low	**531** 75.2%	148 21.4%	27 3.9%	706 33.8%
	Med	156 22.1%	**335** 48.4%	201 29.2%	692 33.2%
	High	25 3.5%	173 25.0%	**490** 71.2%	688. 33%
		712 100%	656 100%	718 100%	2086

Average precision = 65%

Results – Key Takeaways

- ❖ **At the category level, viral load, hospitalizations, and CLI, each have a strong association with cases.**
- ❖ **Forecast performance depends on data or the time frame used for the model training.**
- ❖ **Adding other important signals such as CLI and hospitalizations improves model precision.**
- ❖ **Forecast model performance is similar for cases and hospitalizations.**
- ❖ **In most districts, probability forecasts outperform the Naïve model (uniform).**

Evaluation Metrics

- To evaluate the probability forecasts we use **Brier score**:

$$BS(D,t) = \sum_{i=1}^{3} \left(p_i(D,t) - o_i(D,t)\right)^2$$

here $p_i(D,t) = P(Q_{cases}(D,t) = i)$ and $o_i(D,t)$ is the observed probability, for district D at time t.

- To evaluate the point forecasts $\widehat{Q_{cases}}(D,t) = \arg\max_i p_i(D,t)$, we use the **precision** metric.

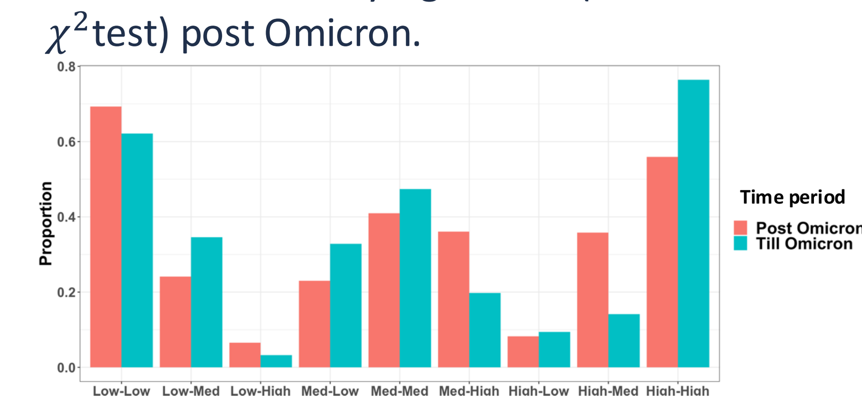
$$Precision(i) = \frac{Number\ of\ correctly\ classified\ forecasts\ in\ category\ i}{Total\ number\ of\ forecasts\ in\ category\ i};\ i \in \{Low, Med, High\}$$

Impact of time-frame used for analysis and forecast models: Before and after Omicron wave

- ➢ **Precision of the forecast model (incorporating all signals) decreases after the Omicron wave.**

	Precision Low	Precision Med	Precision High	Average precision
Till Omicron	89.7%	64.6%	83.2%	79.6%
Post Omicron	76.4%	54.2%	62%	64.3%

- ➢ The association between Cases and Wastewater remains statistically significant (based on χ^2test) post Omicron.



Comparison between Case forecasts & Hospitalization forecasts

We use a multinomial model with all signals:

$$\eta(Q_{cases}(t)) = \beta_0 + \beta_1 Q_{cases}(t-1) + \beta_2 Q_{hosp}(t-1) + \beta_3 Q_{CLI}(t-1) + \beta_4 Q_{VL}(t-1).$$

A similar model is used to model $Q_{hosp}(t)$.

- **Combining forecasts from all districts, the average precision is similar for hospitalizations and cases.**

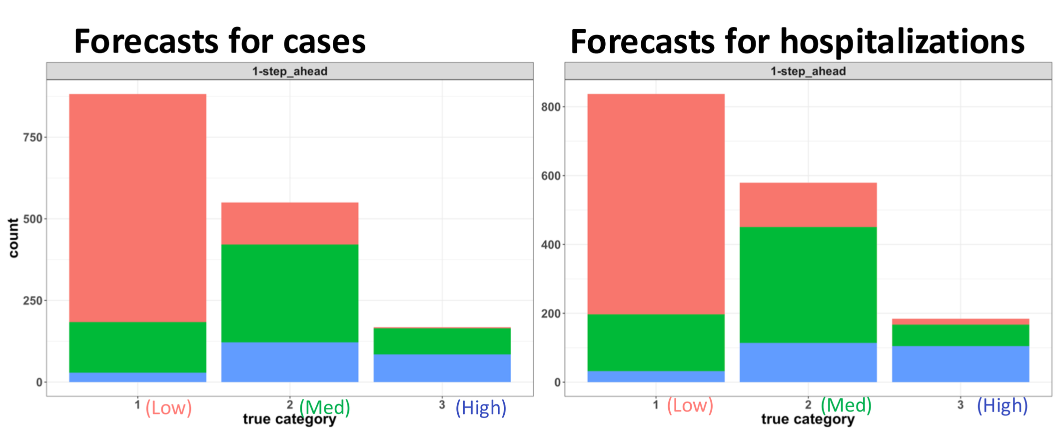


Fig : Overview of the combined district level forecasts. Comparison between 1–week ahead forecast categories and the true categories.

Temporal probabilistic forecasts and their evaluation using Brier score

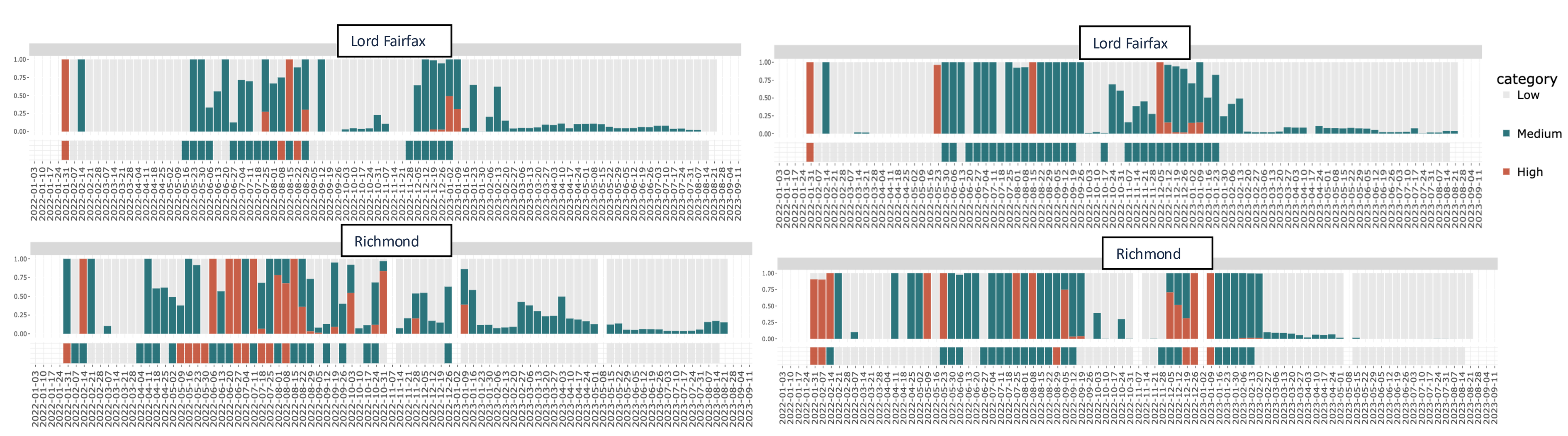


Fig : Probability forecasts and ground truth over time for two districts – Lord Fairfax and Richmond. Left side represents case forecasts and right side represents hospitalization forecasts.

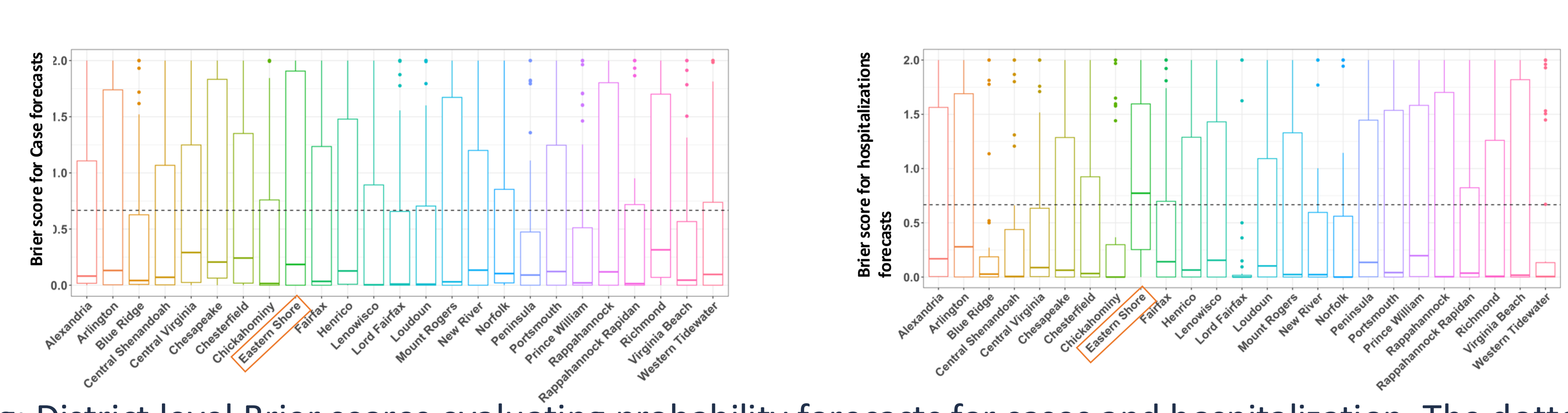


Fig: District level Brier scores evaluating probability forecasts for cases and hospitalization. The dotted line denoting $\frac{2}{3}$, corresponds the Brier Score of the Naïve model assigning equal probability for each category.

Observation & Challenges:

- Quantile categories lack robustness: highly sensitive to new values at any time
- Variable performance across districts due to varying data quality.

Fig: Forecast categories vs true categories for Eastern Shore.

Funding Acknowledgments:

CSTE/CDC 5 NU38OT000297, NIH Grant 1R01GM109718, NSF BIG DATA Grant IIS-1633028, NSF Grant No.: OAC-1916805, NSF Expeditions in Computing Grant CCF-1918656, CCF-1917819, NSF RAPID CNS-2028004, NSF RAPID OAC-2027541, US-CDC 75D30119C05935s, VDH COVID-19 Response Contract VDH-23-501-0253, CDC Cooperative Agreement CK22-2204

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE