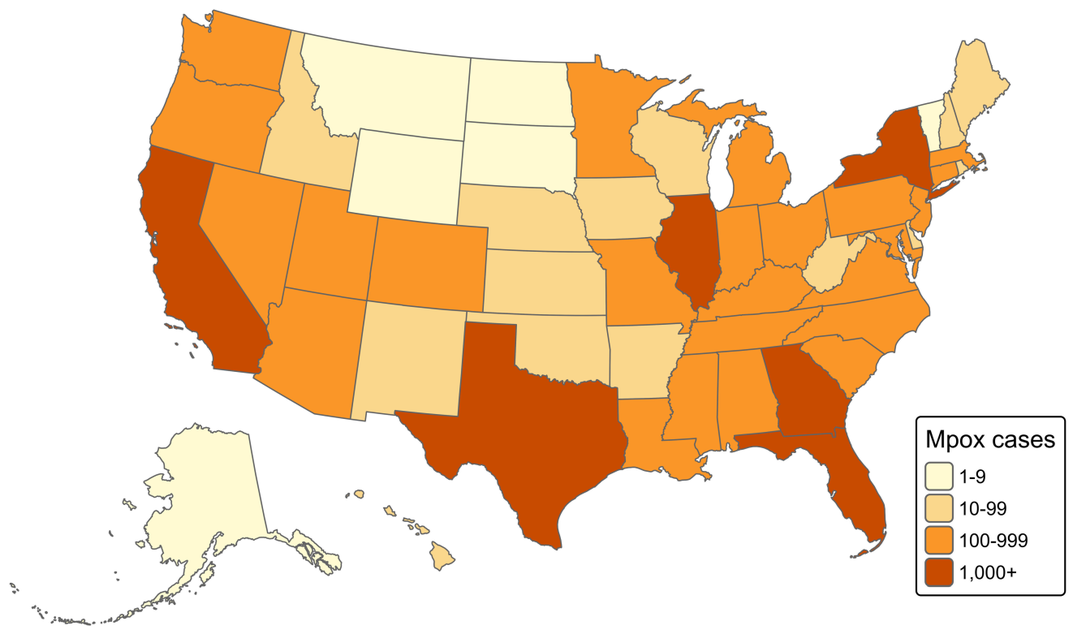
From Clicks to Cases

The predictive power of Wikipedia pageviews for mpox cases

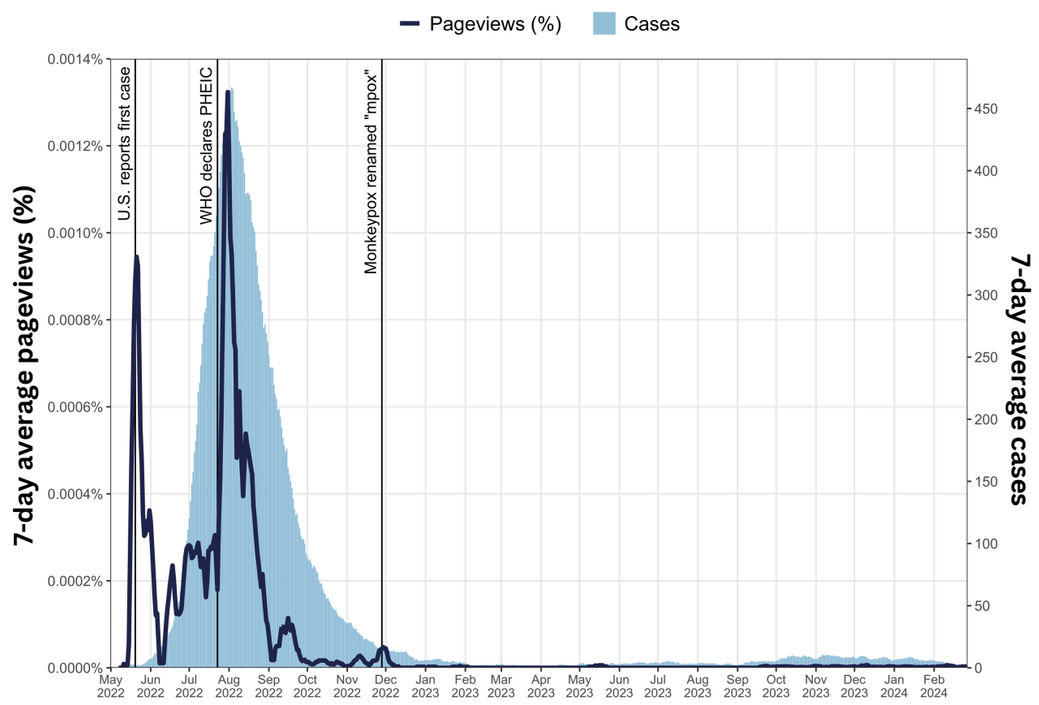
Background

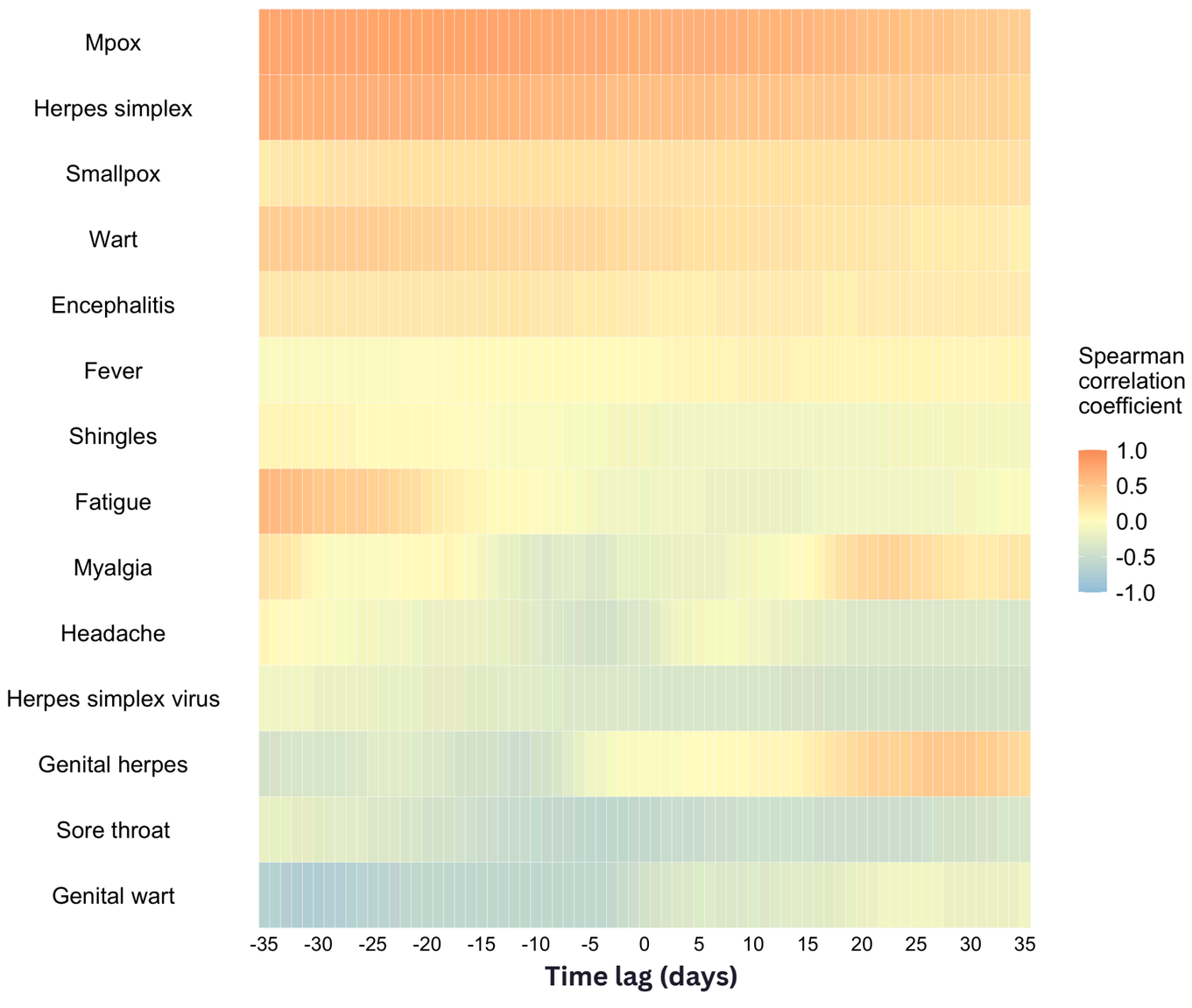
As the world becomes increasingly interconnected and climate change elevates zoonotic spillover risks, global outbreaks become more likely. Traditional surveillance methods, though accurate, often face delays. By contrast, internet activity data can provide real-time tracking of health-related information-seeking behaviors. Wikipedia's openly available and detailed pageview data shows promise for outbreak detection. This paper explores the predictive value of anonymized country-level Wikipedia pageview data for forecasting case incidence, using the 2022-2024 mpox (formerly monkeypox) outbreak in the United States as a case study.



Methods

This study uses data on daily mpox cases and Wikipedia pageviews for mpox-related articles from May 2022 to February 2024. First, a lag analysis evaluates correlation between mpox cases and pageviews for various time lags ranging between -35 to +35 days. Next, impulse response and Granger-causality tests using Vector autoregression (VAR) are applied to explore predictive relationships between mpox cases and pageviews. Finally, multivariate linear regression is used to model the relationship between pageviews and mpox incidence, assessing four models for predictive accuracy.





Results

1. The lag analysis revealed statistically significant correlations between pageviews of several mpox-related articles and case incidence, with relationships varying by lag.
2. The VAR model demonstrated significant predictive power between Wikipedia pageviews and mpox cases, validated by high R-squared values and impulse response functions. Furthermore, Granger-causality tests confirmed predictive relationships but no instantaneous causality was established.
3. The four predictive models showed varied results, with the simplest models underestimating declines, while more complex models faced issues such as overfitting, indicating the need for further work to achieve optimal performance.

Conclusion

Public health officials should integrate digital epidemiological surveillance methods into early warning systems to effectively mitigate future outbreaks. To enhance data availability without compromising privacy, the Wikimedia Foundation should consider population size and internet usage rates when setting data release thresholds. Aggregating pageviews by Wikidata ID could further protect language minorities while increasing data accuracy.

Steven Kerr | Master of Data Science for Public Policy – Class of 2024 | smkerr.github.io | Hertie School