Sara Kim

Klamon/Winston Award
Ph.D. Student, Epidemiology
First Year ARCS Scholar

EMORY UNIVERSITY

Modern tools for modern threats: Innovative application of machine learning methods to infectious disease research

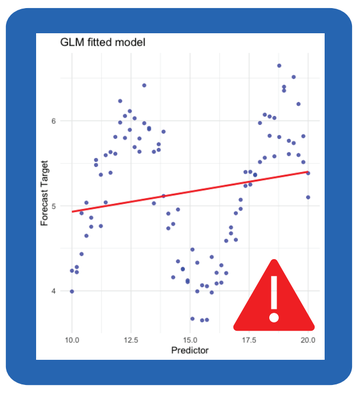
Disease forecasting is an important public health tool that anticipates the timing and severity of surges in disease activity, enabling public health agencies to prepare for and rapidly respond to outbreaks. This research uses machine learning to develop forecasts for norovirus outbreaks in the United States.

BACKGROUND

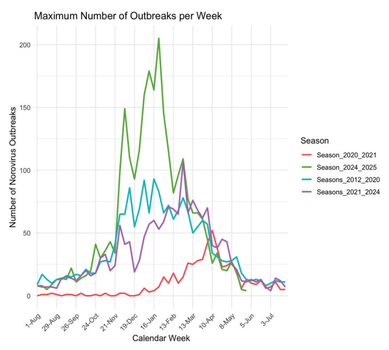
Disease forecasting anticipates **timing** and **severity** of potential surges in disease

Traditional forecasting methods such as generalized linear models (GLMs) make **rigid assumptions** that often **do not reflect real-world data**.

Machine learning may offer improved accuracy



Research Aim: Evaluate traditional versus machine learning methods to improve forecasting norovirus outbreaks in the United States



Norovirus is hard to predict due to its **complex epidemiology** and frequent **underreporting**, but this makes it good candidate to try machine learning on

- **Timely:** Highest surge in norovirus outbreaks in the U.S. this past winter
- **Epidemiologically complex:** genetically diverse, imperfect seasonality, many modes of transmission
- **Incomplete information:** generally underreported disease, understudied for forecasting
- **High burden:** 19-21 million illnesses, 400,000 emergency department visits, and 56,000-71,000 hospitalizations per year

6

METHODS

Generalized linear model (GLM)	Least absolute shrinkage and selection operator (LASSO)	Ensemble machine learning
Advantages • Easy implementation • Interpretable Challenges • Vulnerable to misspecification • Cannot handle many predictors • Cannot capture non-linear relationships	Advantages • Automatic predictor selection • Handles many predictors • Interpretable Challenges • Cannot capture non-linear relationships	Advantages • Can capture non-linear relationships • Can include automatic predictor selection • Can handle many predictors Challenges • Blackbox

Data source: Centers for Disease Control and Prevention's National Outbreak Reporting System
- Training set: 2009-2018
- Testing set: 2018-2022

Forecast methods:
- Generalized linear model
- Regularized regression using LASSO[6,7]
- Ensemble model using SuperLearner[8]

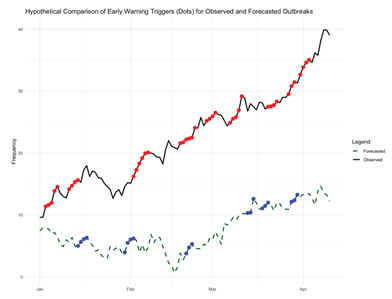
Forecast target:
- 7-day rolling average of norovirus outbreaks nationally
- 1, 7, 14, and 28-days ahead

Method evaluation:
- Mean squared error (MSE)
- Weighted interval score (WIS)

EXPECTED RESULTS

Left: Heatmap to compare quantitative performance across methods and data combinations
Right: Comparison of early warning alerts across observed and forecasted norovirus outbreaks

	Best GLM	Ensemble	LASSO
Lagged terms alone	WIS	WIS	WIS
Add syndromic surveillance	WIS	WIS	WIS
Add laboratory data	WIS	WIS	WIS
Add syndromic surveillance and laboratory data	WIS	WIS	WIS



PUBLIC HEALTH IMPACT

Systematic and comprehensive evaluation of traditional versus machine learning-based methods using enhanced data sources

- Develop norovirus forecasting infrastructure
- Improve preparedness and response
- Protect vulnerable populations
- Empower public health agencies
- Scale advanced methods to study other pathogens

1

Scholar Awards Celebration
November 12, 2025

SCHOLAR AWARDS CELEBRATION ARCS FOUNDATION ATLANTA

Igniting Innovation in Georgia