

# Connecting Information to Improve Detection in Event- Based Surveillance



David Buckeridge, MD PhD  
Professor and Canada Research Chair  
McGill University, Montreal, Canada

# Context

## EPI-AI project

- **Goal:** incorporate modern NLP into EBS and develop aberration detection methods that exploit modern NLP
- Using BioCaster as model system


## Study of global variation detection in EBS

- **Goal:** Assess variation in frequency-based aberration detection using seasonal influenza as a proxy
- Data from HealthMap and EIOS

## GPHIN workflow analysis

- **Goal:** Develop user experience for incorporating clustering of reports into EBS
- User stories, interface mock-ups, and algorithm development with GPHIN analysts

### Bioinformatics




Article Navigation

JOURNAL ARTICLE

#### BioCaster in 2021: automatic disease outbreaks detection from global news media

Zaiqiao Meng, Anya Okhmatovskaia, Maxime Polleri, Yannan Shen, Guido Powell, Zihao Fu, Iris Ganser, Meiru Zhang, Nicholas B King, David Buckeridge, Nigel Collier

Bioinformatics, Volume 38, Issue 18, 15 September 2022, Pages 4446–4448,  
<https://doi.org/10.1093/bioinformatics/btac497>



**JMIR Publications**  
Advancing Digital Health & Open Science

Articles 0

JMIR Public Health and Surveillance

Published on 31.10.2022 in Vol 8, No 10 (2022): October

Preprints (earlier versions) of this paper are available at <https://preprints.jmir.org/preprint/36211>, first published January 05, 2022.

#### Global Variations in Event-Based Surveillance for Disease Outbreak Detection: Time Series Analysis

Iris Ganser<sup>1,2</sup>; Rodolphe Thiébaud<sup>2,3</sup>; David L Buckeridge<sup>1</sup>

1. What *influences detection* in event-based surveillance?

2. How can we *exploit these factors* to improve detection?

1. What *influences detection* in event-based surveillance?

2. How can we *exploit these factors* to improve detection?

# Evidence About Determinants of Detection

## Reports

- Source
- Frequency
- Extraction



## Disease

- Novelty
- Presentation

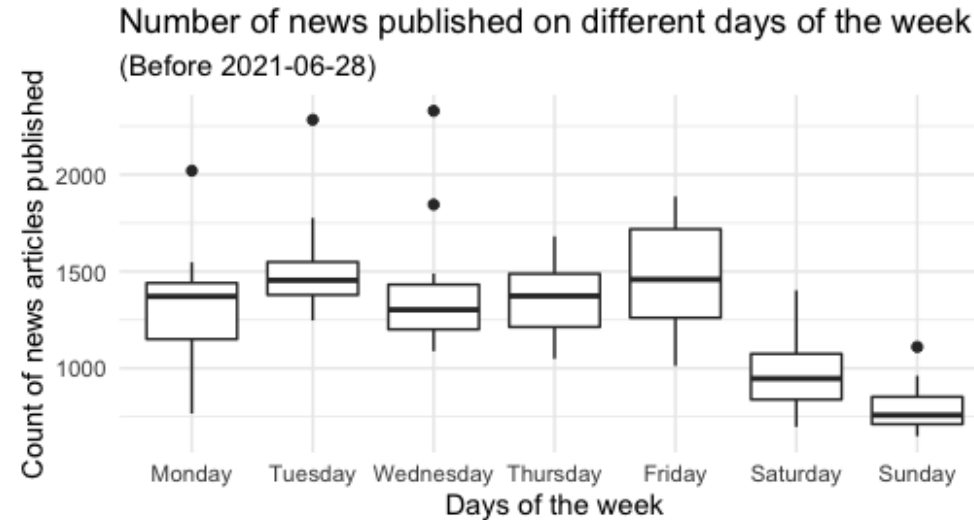


## Context

- Health system
- Media landscape
- Resources

# Reports

- Source
  - Type of media
  - Formal vs informal
  - Language
- Frequency
  - Day-of-week
  - Over time within an event
  - Crowding by other disease
- Extraction of information

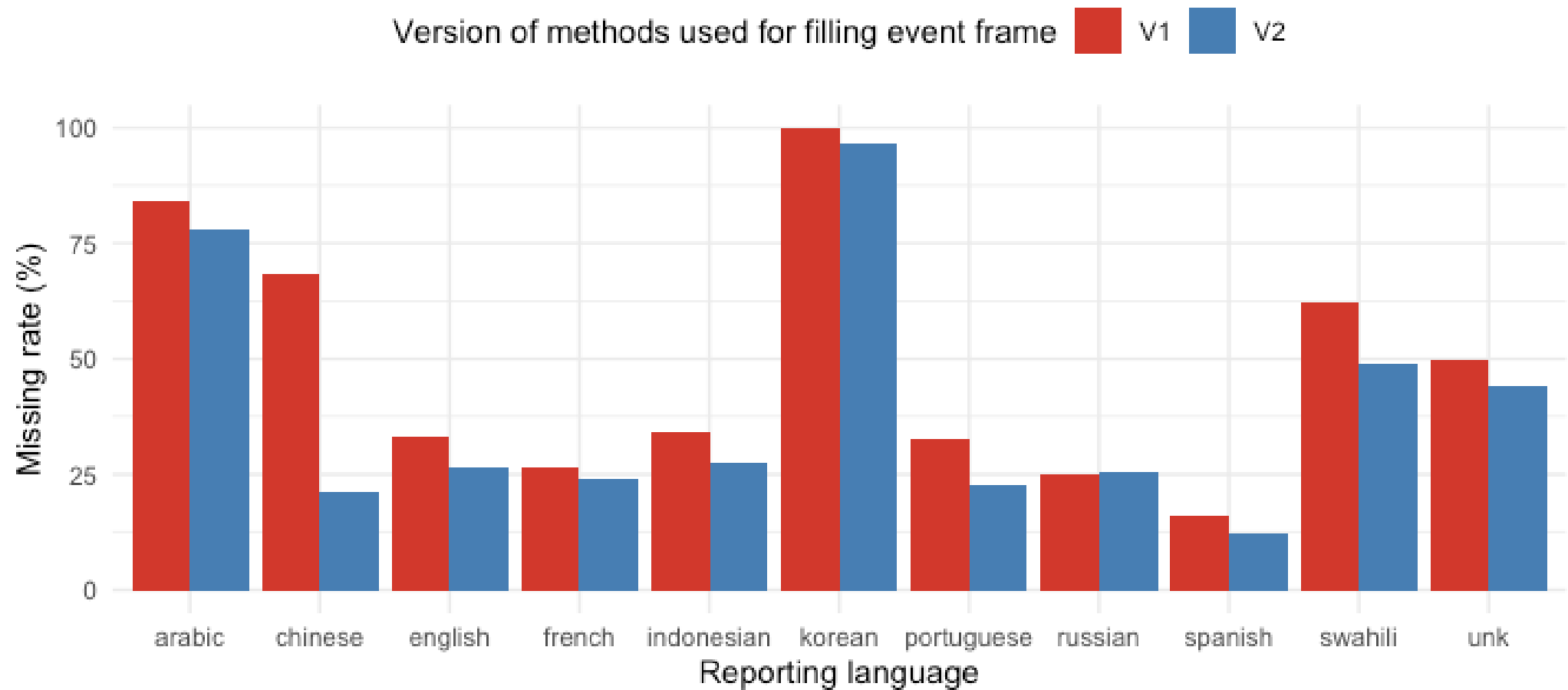


Scales D, Zelenev A, Brownstein JS. Quantifying the effect of media limitations on outbreak data in a global online web-crawling epidemic intelligence system, 2008-2011. *Emerg Health Threats J.* 2013 Nov 8;6:21621.

Smith KC et al. Understanding newsworthiness of an emerging pandemic: international newspaper coverage of the H1N1 outbreak. *Influenza Other Respir Viruses.* 2013 Sep;7(5):847-53.

Drivers of Emerging Infectious Disease Events as a Framework for Digital Detection. Olson SH et al. *Emerg Infect Dis.* 2015 Aug;21(8):1285-92.

# Variation in Extraction by Report Language

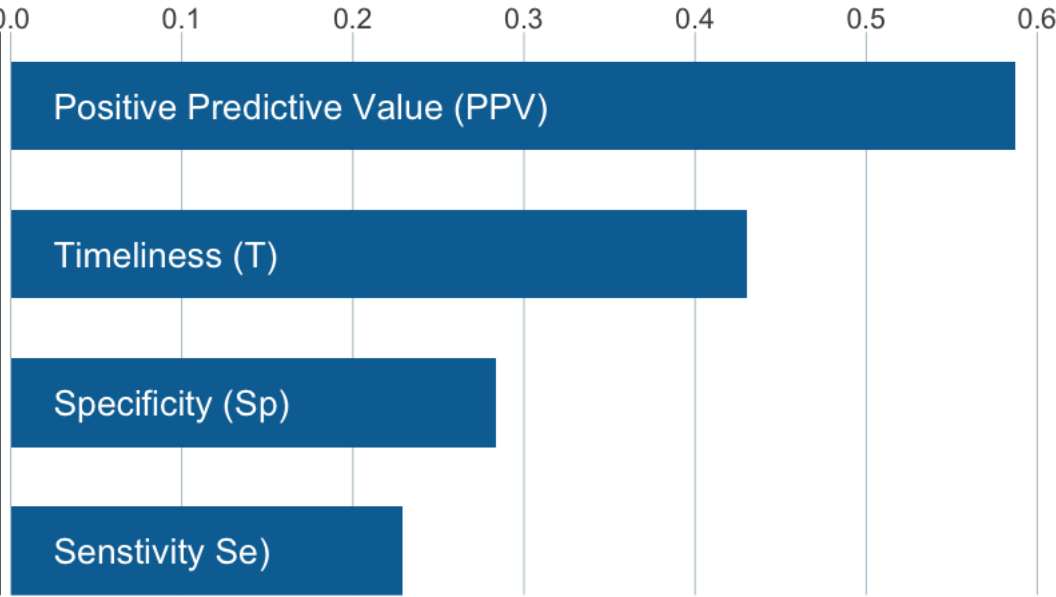


# Context



## Influence of Contextual Factors on Aberration Detection

Proportion of Variance Explained



## Influence of Individual Factors

Significance of Factor for Predicting Metric

	PPV	T	Se	Sp
<b>Geography</b>				
Latitude (+)				
Tropical (vs N Temp)				
<b>Other Factors</b>				
Press Freedom Index (+)				
Not English (-)				



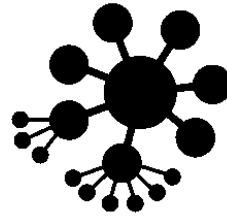
1. What *influences detection* in event-based surveillance?

2. How can we *exploit these factors* to improve detection?

# Strategies to Improve Detection



Account for  
background variation  
in frequency of reports



Connect information  
about same event  
across reports



Flag known biases  
where there is no  
technical solution

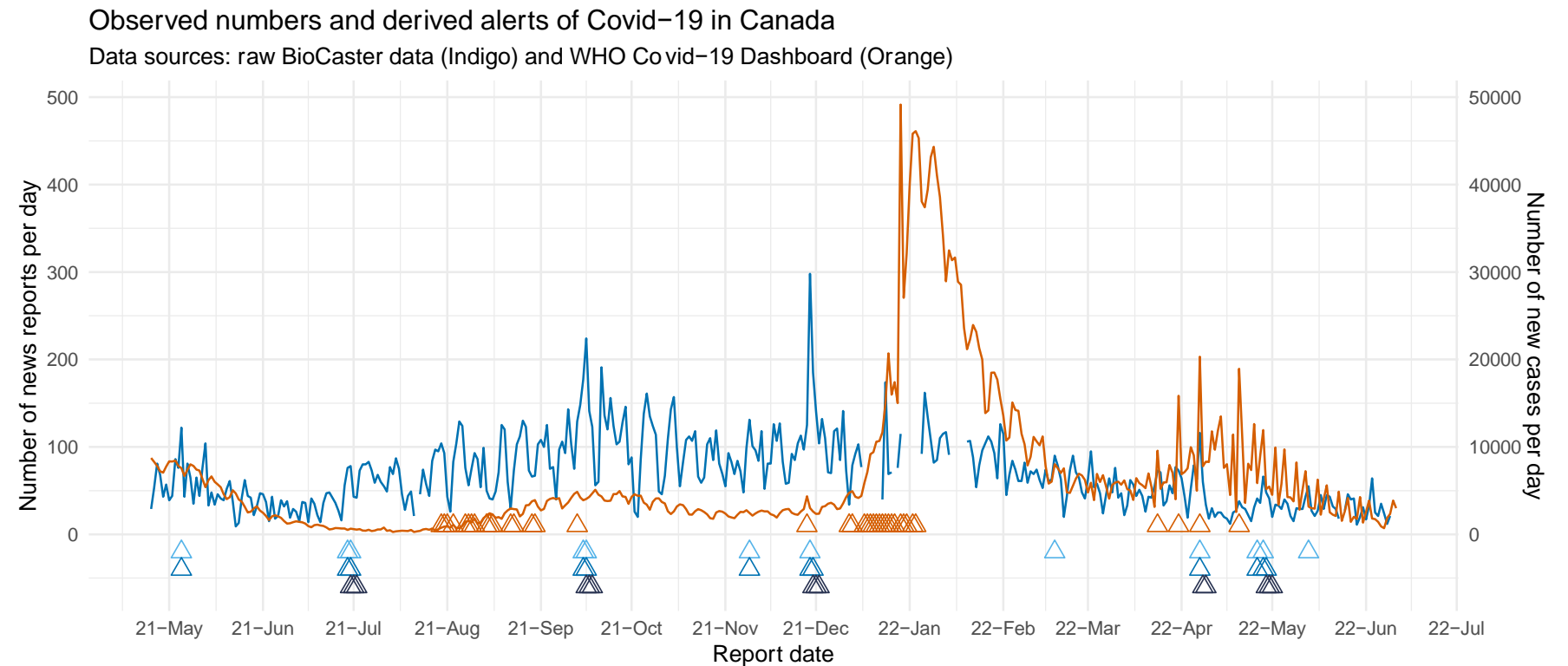
# Account for Background Variation in Frequency of Reports



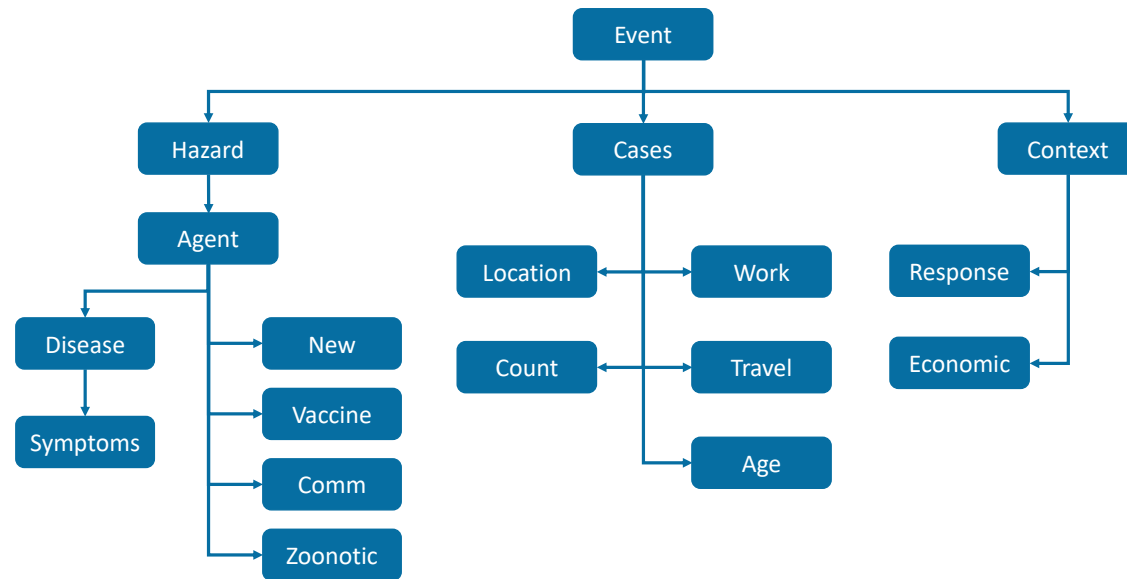
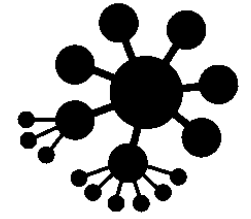
Source  
variation

Day-of-week

Seasonal



# Connect Information Across Reports



Information model (ontology) for an event

The screenshot shows a web application interface for monitoring events. The main menu includes "All languages", "All articles", and "Saved events". The "Saved events" section displays a map and timeline showing the spatiotemporal distribution of reports within the cluster. The "Event summary" section provides details about the event, including location, disease, cases, deaths, symptoms, age, population, and risk. The "Articles" section lists selected original articles used to generate the event summary. Annotations explain the following features:

- Saved events are accessed through main menu**: Points to the "Saved events" tab in the main menu.
- Multiple events can be monitored at the same time. The user can examine them individually**: Points to the "Unknown disease, DRC" and "Lassa fever, Nigeria" links.
- Original articles selected for this event and used to generate event summary**: Points to the "Articles" section.
- Map and timeline show spatiotemporal distribution of reports within the cluster**: Points to the map and timeline visualization.
- Event summary shows information synthesized from all selected reports available so far. Specific fields will have to be identified based on the user needs and feasibility of automatic extraction relevant information**: Points to the "Event summary" section.
- Clicking on any field in the event summary highlights parts of the text in the original reports from which the displayed information was synthesized**: Points to the "Symptoms" field in the event summary.
- Clustering algorithm quantifies the similarity between the saved group of reports above and other articles. Guided by these similarity scores, the user can easily add more articles to the cluster**: Points to the "Find similar articles" section.

Clustering reports and communicating information

# Flag Potential Biases for Users



- Technical solutions are only available for some of the known factors that influence detection
- Informing EBS users of the unaddressed factors could limit the negative effect of these factors on detection
- Machine translation as an example
  - Accuracy of machine translation varies by language (as does NLP in general)
    - Influences completeness of information extraction
    - Influences available information and detection by geographical region
  - Accuracy for each language can be measured and communicated to users in the context of a report of event

# Summary

Multiple factors influence aberration detection in EBS



Reports



Disease

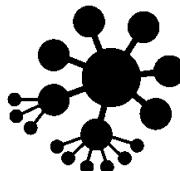


Context

These factors can be exploited to improve detection



Adjust for Background



Connect Reports



Flag Biases

# Acknowledgements

- EPI-AI Project
  - Nigel Collier (Cambridge), Nicholas King (McGill), Maxime Polleri (McGill), Zaiqiao Meng (Cambridge), Yannan Shen (PhD student), Meiru Zhang (Cambridge), Zihao Fu (Cambridge)
- Global Variation in Detection
  - Iris Ganser (McGill), Rodolphe Thiebault (Bordeaux), HealthMap, EIOS
- Clustering in GPHIN
  - Anya Okhmatovskaia, Yannan Shen, Erin Rees, Victoria Ng, GPHIN