Infectious Disease Patterns in Global Online Media Data: Detection, Reasoning, and Evaluation

David Buckeridge, MD PhD FRCPC
Professor | School of Population and Global Health, McGill University
CIHR Chair | E-Health Interventions
Director | RI-MUHC Data Warehouse
david.buckeridge@mcgill.ca

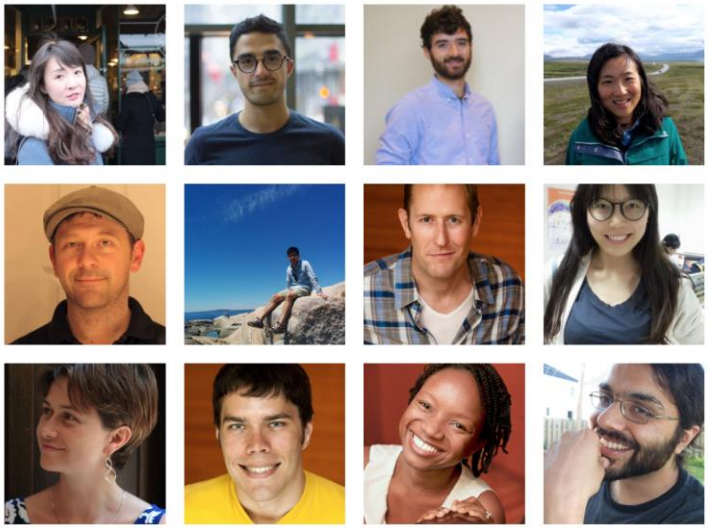
McGill

World Health Organization (WHO), Epidemic Intelligence from Open Sources (EIOS) Global Technical Meeting
November 12-14, 2019, Seoul, Korea

The Surveillance Lab

ABOUT MCHI ←

The Surveillance lab within the Clinical and Health Informatics Research Group at McGill University brings together a vibrant multidisciplinary team of over 20 investigators, public health practitioners, clinicians, research staff, students and software developers, all dedicated to conducting research and development of computational methods and software that has immediate impact on improving population health through the science and practice of biosurveillance. The Surveillance Lab is funded by several sources including the Canadian Foundation for Innovation, the Canadian Institutes of Health Research, a Canada Research Chair, the Bill and Melinda Gates Foundation, the National Sciences and Engineering Research Council, the Centers for Disease Control and Prevention and many other sources. On many projects, we work closely with public health practitioners in Quebec and from around the world. The computerized solutions we have developed are used by public health agencies in Quebec, Canada, and internationally.



Detection

Reasoning

Evaluation

Detecting Events among Reports

‘Connecting-the-dots’ by integrating relevant concepts from across reports remains a significant problem [G7 initiative of the Global Health Security Action Group (GHSAG)]

One task is to connect media reports from different sources that refer to the same event

Needs for “Connecting Dots”

Common “Language” across reports from systems

- Extracted entities
- Free text (machine translated)

Methods for merging reports

- Distance metric (semantic, quantitative)
- Unsupervised clustering

Unambiguous Concept Representation

"Exploiting the promise of EIOS will depend on our ability to align data across reports and systems with comparable and consistent formats and contextual meaning."

Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. N Engl J Med. 2018 Oct 11;379(15):1452-1462.

Degrees of Meaning (Semantics)

Terminology or Lexicon

the set of terms used in a domain or language

Concepts or terms

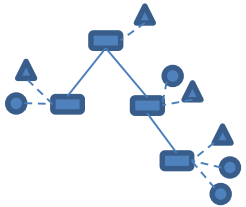
Controlled Vocabulary

set of preferred terms for a domain

Set of preferred terms with synonyms

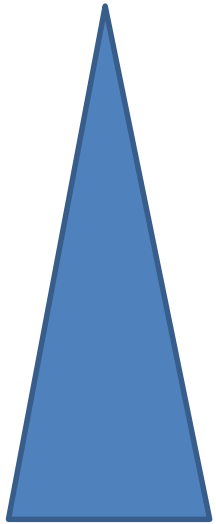
Ontology

controlled vocabulary with formal semantic relationships



Taxonomic and other associations between terms

Uses of Formal Semantic Models



Searching heterogeneous data

Exchanging data among applications

Natural language processing

Integrating information

Encyclopedic representation of knowledge

Computer reasoning with data

Rubin, D. L., Shah, N. H., & Noy, N. F. (2008). Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, *9*(1), 75–90.

Detection

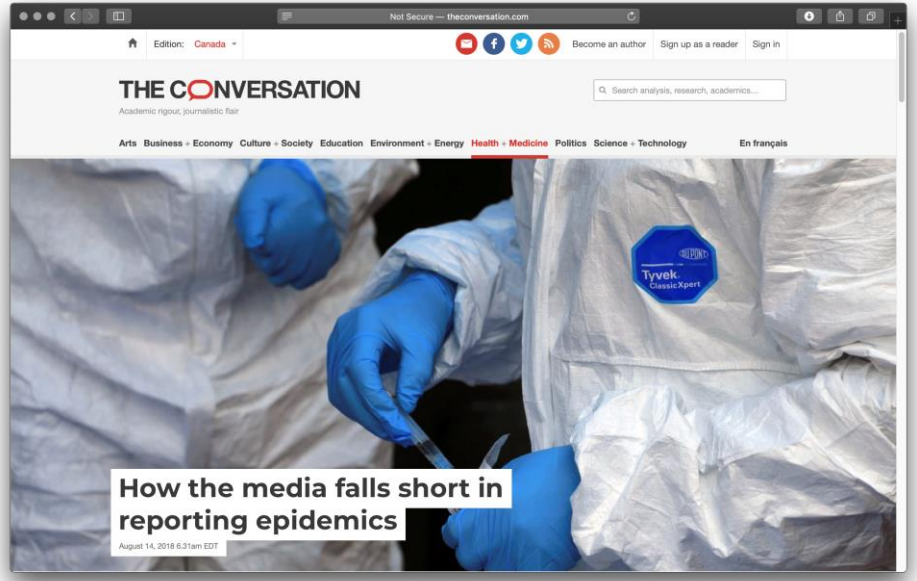
Reasoning

Evaluation

Making Sense of Detected Events

Interpreting new information from detected event

Possible biases in reporting (with Nicholas King, McGill)



Interpreting New Information

Task is to understand how an event changes risk assessment

Challenges include

- New information may be qualitative, uncertain
- Prior information on risk may be incomplete, low quality
- Other contextual information may be relevant (e.g., media bias)
- A mechanism is needed to update prior risk, accounting for event and context

Methods for updating risk assessment

- Automated reasoning (i.e., using encoded knowledge to interpret data)
- Bayesian hierarchical modeling

Effects, Sources of Media Bias

Differences in media across countries can

- Affect reporting accuracy, sensitivity, specificity
- Create or amplify heath inequalities

Reporting may be influenced by many factors

- Geography, Time
- Media Penetration, Media Economy
- Disease, Social Advantage

Plan is to systematically define, identify, map biases

Detection

Reasoning

Evaluation

Iris Ganser (Alexandra Schmidt, Rodolphe Thiebaut)
Evaluation of event-based internet biosurveillance systems for determination of seasonal influenza onset

Evaluation Project

Objective: To evaluate global variation in disease outbreak detection from online media reports

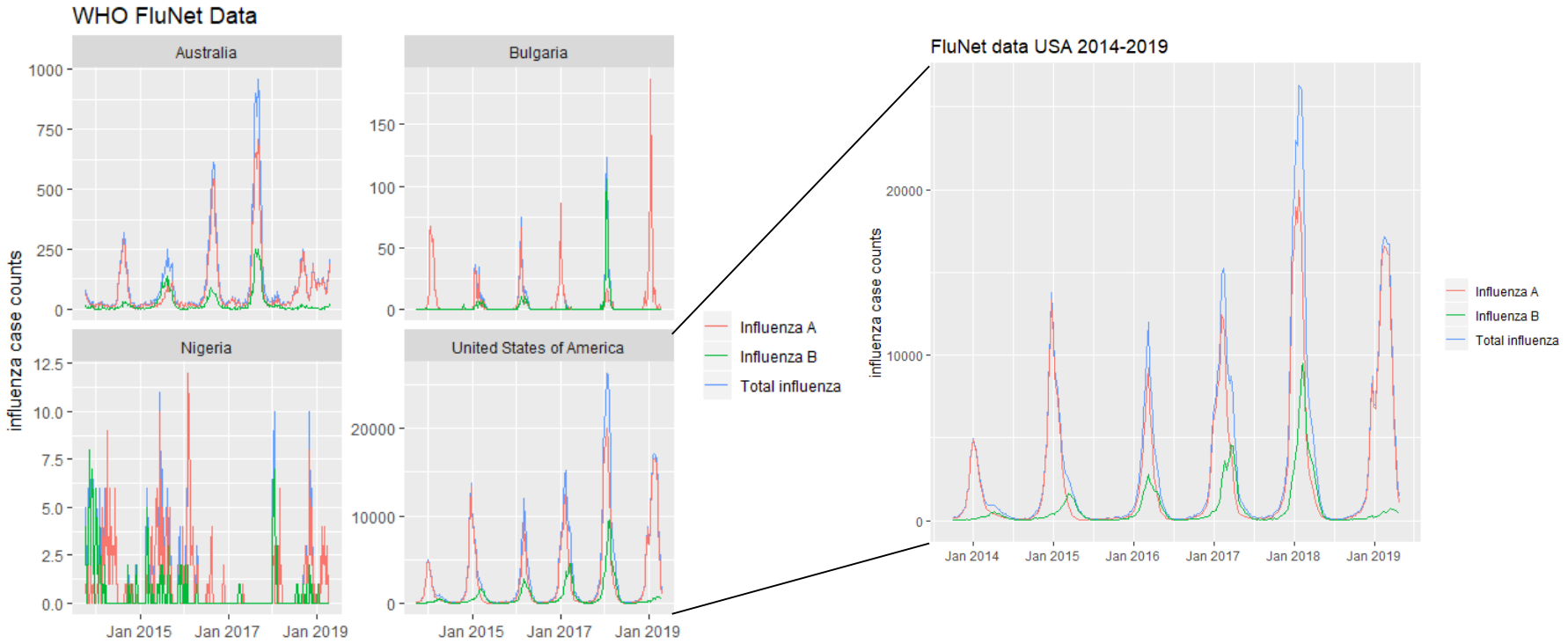
Data sources: Online media reports (HealthMap, EIOS), Laboratory (FluNet – “gold standard”)

Model organism: Human seasonal influenza (proxy)

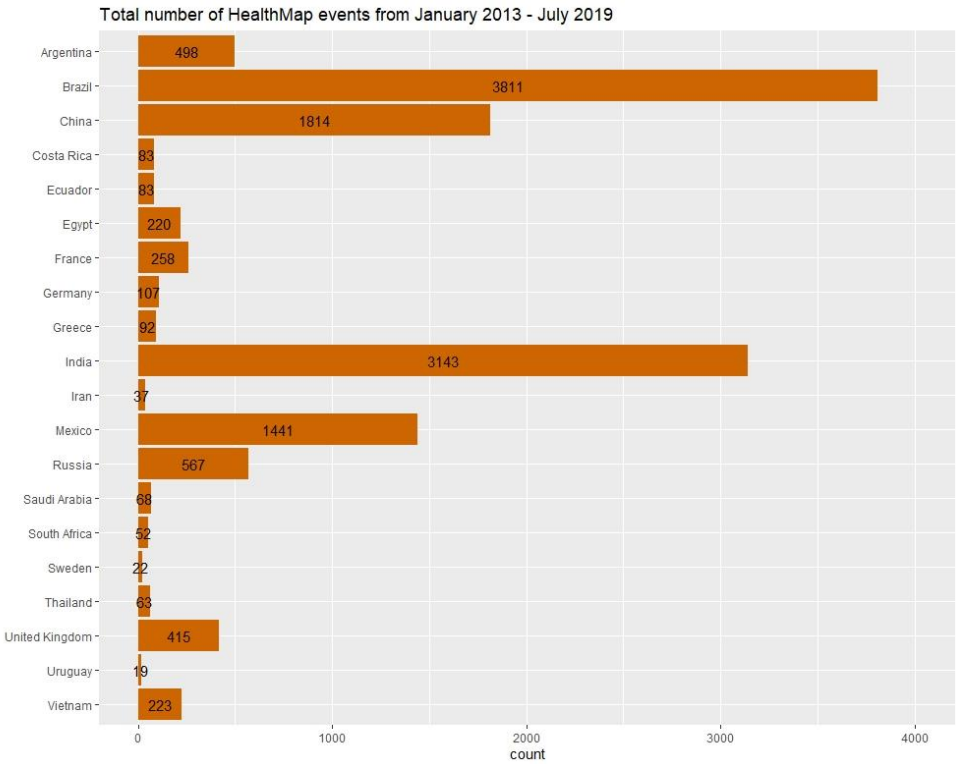
Methods: Bayesian dynamic linear models (epidemic onset and curve)

Scope: 24 countries across all Influenza regions

Initial results - WHO FluNet data

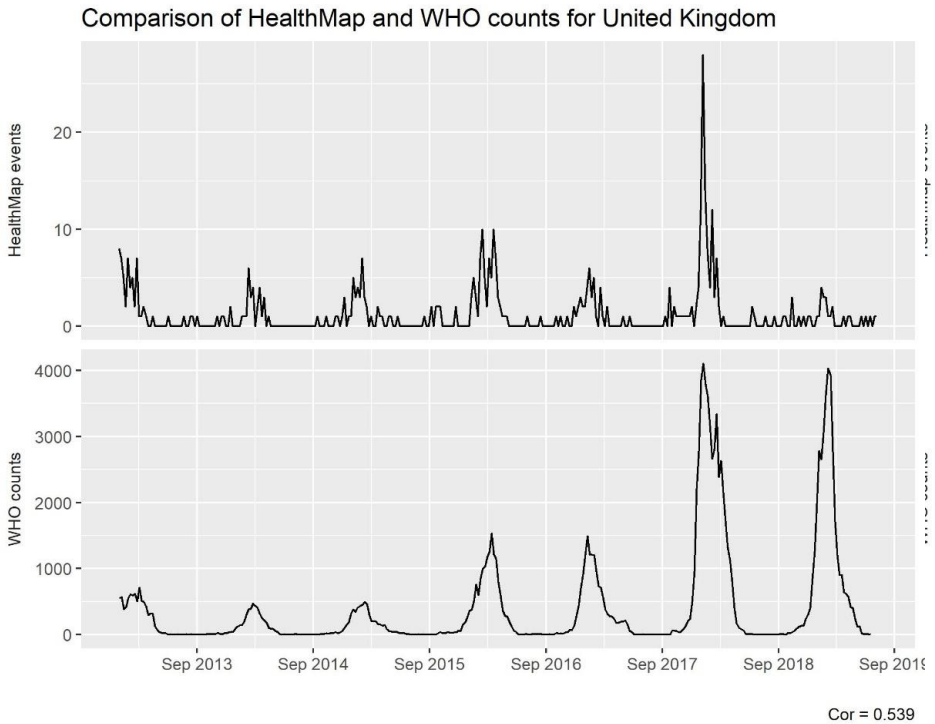


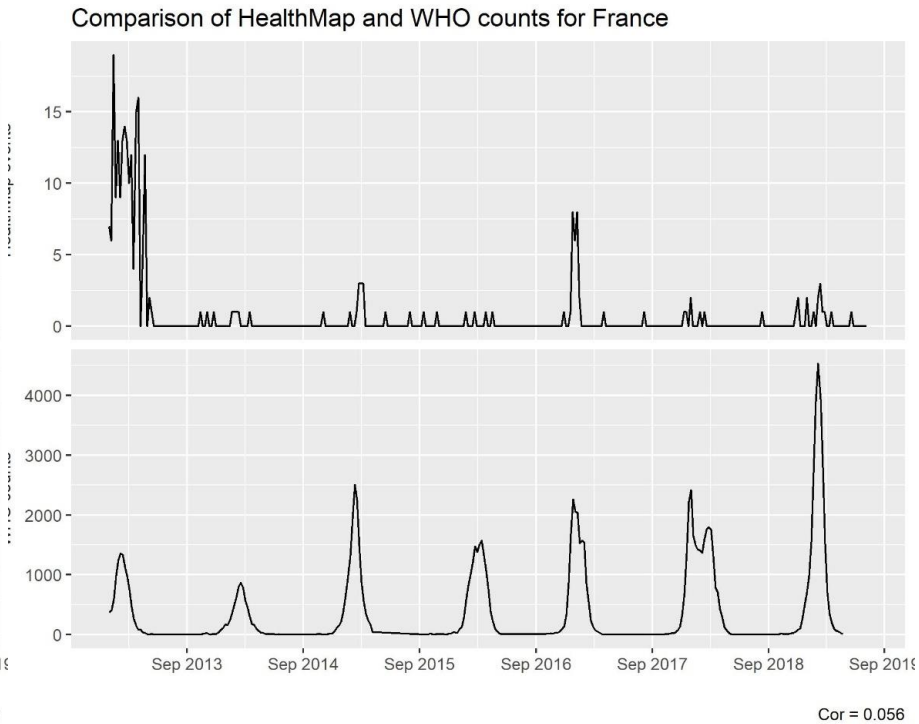
Initial Results - HealthMap Reports by Country



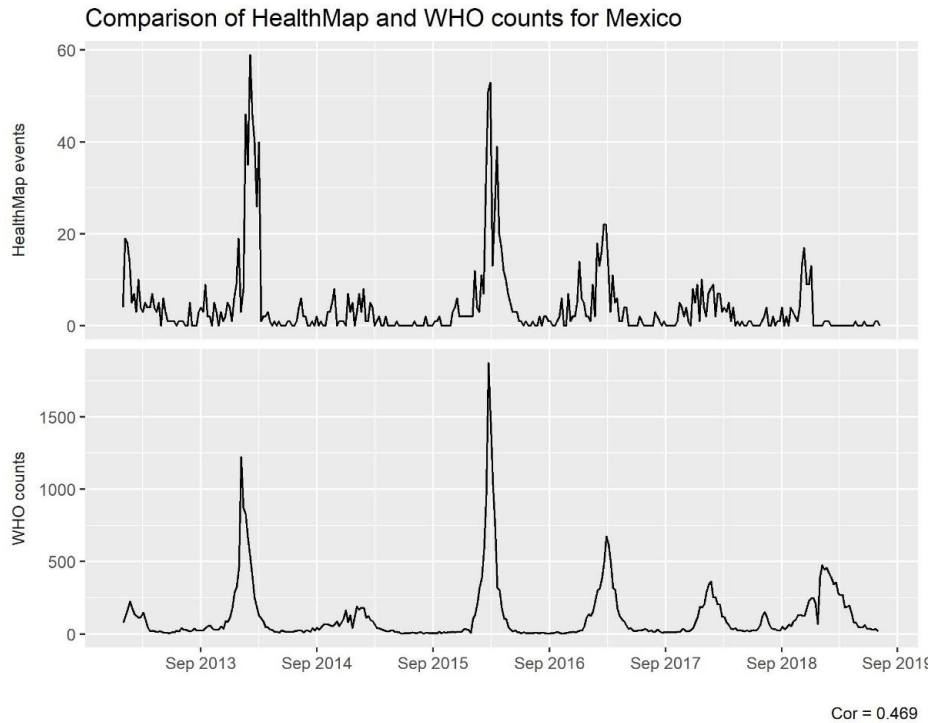
- **Total number of reports varies strongly by country and language**
- USA has even considerably more events we have not received the data for the whole time period yet

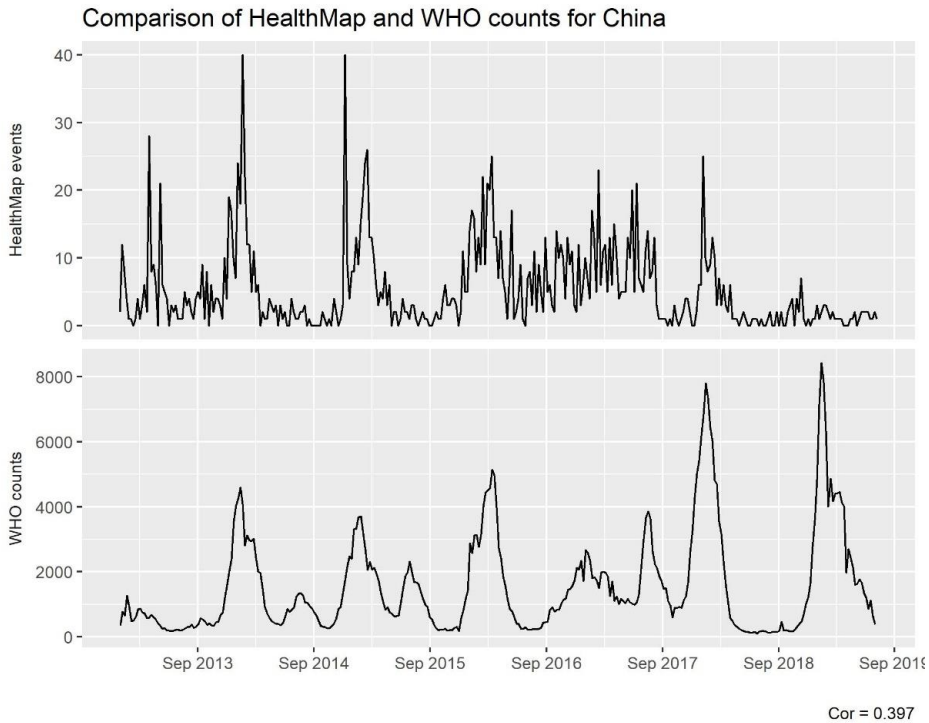
Initial Results – UK and France





Initial Results – Mexico and China





Next Steps for Research

Detection and Reasoning (Pending funding)

- Identify requirements for knowledge representation
- Establish EIOS test data set for ‘connecting dots‘
- Assess potential of symbolic and Bayesian updating methods
- Develop review strategy for characterizing media biases

Evaluation Study (Spring 2020)

- Filter, extract, incorporate EIOS data
- Refine modeling strategy (model selection, time lag for FluNet)
- Inclusion of other data sources