Evaluation of event-based internet biosurveillance for multi-regional detection of seasonal influenza onset

Iris Ganser

2nd year of the Master Sciences, Technologies, Santé:

Santé Publique - Parcours Public Health Data Science

Evaluation of event-based internet biosurveillance for multi-regional detection of seasonal influenza onset

A thesis presented from

The surveillance lab, McGill University

to

The Digital Public Health Graduate Program, University of Bordeaux

Submitted on June 3rd, 2020

By Iris Ganser

Supervisors:
Dr. David Buckeridge, Professor
Dr. Rodolphe Thiébaut, Professor

Funding by PIA3 (Investment for the future)
supporting the EUR Digital Public Health Graduate program

Abstract

Infectious diseases remain a major public health problem worldwide. Hence, event-based surveillance (EBS) was developed to allow more timely detection of infectious disease outbreaks using web-based data, but EBS systems have never been evaluated on a global scale. Thus, the objective of this thesis is to evaluate the ability of EBS to detect epidemic outbreaks of influenza in 24 countries worldwide. Additionally, factors influencing system performance will be identified. Data were obtained from two EBS systems, HealthMap and EIOS. Publicly available weekly virological influenza data were gathered from the FluNet platform as gold standard data. Bayesian change point analysis was used to detect the beginning and end of influenza epidemics. Then, evaluation metrics were calculated, with timely sensitivity, i.e. outbreak detection within the first two weeks after onset, as the main outcome of interest. System performance varied widely between countries and systems, and timely detection of outbreaks was poor in both systems, with HealthMap showing consistently better performance than EIOS. Whereas data abundance influenced the performance of both systems, the human development index (HDI) was influential for HealthMap, and EIOS performance was dependent on a country's geographical location. It could be shown that application of biosurveillance methods to the frequency of online media reports about influenza from two EBS was not able to detect seasonal influenza outbreaks in a timely manner. However, extraction and analysis of additional information from online media and the integration of EBS with other data sources may help to attain this goal.

Keywords: event-based surveillance, influenza, internet, public health surveillance, Bayesian change point analysis

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. David Buckeridge and Dr. Rodolphe Thiébaut for their valuable input and continuous guidance during my thesis. I wish to thank Dr. Buckeridge for providing his public health surveillance expertise, excellent insights, and consistent support during this process. I would also like to express my gratitude for his financial support through a stipend. I would like to thank Dr. Thiébaut for his great motivational encouragement, his continuous feedback and his support especially with statistics. I am very thankful to both for supporting my double degree.

My thanks go to all members and the administrative staff of the surveillance lab for creating a stimulating and inclusive working atmosphere. Thanks for all the snacks and the chats!

I would like to thank Dr. Clark Freifeld from HealthMap and Johannes Schnitzler from EIOS for sharing the data from their respective systems for this project with me.

Thanks to the EBOH department at McGill and DPH team at the University of Bordeaux for making the double degree possible. It was such a great experience for my personal as well as my professional development, and I am very grateful for this opportunity.

I would also like to thank my friends for their unwavering support, for all the laughs, good talks and distractions, whether we are physically close to each other or not.

Finally, I would like to express my gratitude to my family for their unconditional support of whichever goals I pursue and wherever I choose to go. Without you, none of this would have been possible.

Table of Contents

List of abbreviations

AIC	Akaike Information Criterion
BCP	Bayesian change point analysis
bcp	R package for Bayesian change point analysis
CCDSS	Canadian Chronic Disease Surveillance System
CDC	Centers for Disease Control and Prevention
COVID-19	Coronavirus Disease of 2019
csv	Comma-separated values
DENV	Dengue Virus
EARS	Early Aberration Reporting System
EBS	Event-based surveillance
EI	Epidemic Intelligence
EIOS	Epidemic Intelligence from Open Sources
EWMA	Exponentially Weighted Moving Average
FAR	False Alarm Rate
GFT	Google Flu Trends
GISRS	Global Influenza Surveillance and Response System
GPHIN	Global Public Health Intelligence Network
HA	Hemagglutinin
HDI	Human Development Index
IBS	Indicator-based surveillance
ICD	International Classification of Diseases
IHR	International Health Regulations
ILI	Influenza-like Illness
LASSO	Least Absolute Shrinkage And Selection Operator
MCMC	Markov Chain Monte Carlo
MediSys	Medial Information System
MERS	Middle East Respiratory Syndrome
ML	Machine Learning
MRSA	Methicillin-resistant *Staphylococcus aureus*
NA	Neuraminidase
NTSS	National Tuberculosis Surveillance System
PFI	Press Freedom Index
PPV	Positive Predictive Value
ProMed	Program for Monitoring Emerging Diseases
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SARS	Severe Acute Respiratory Syndrome
TIU	Total Internet Users
UK	United Kingdom of Great Britain and Northern Ireland
US/USA	United States of America
WHO	World Health Organization

1 Introduction

1.1 Infectious disease surveillance and biosurveillance systems

1.1.1 Why do we need infectious disease surveillance?

Even in the 21st century, infectious diseases continue to threaten populations worldwide. Newly emerging and re-emerging pathogens, microbial drug resistance, and increased opportunities for pathogens to spread through demographic explosion, massive urbanization, and growing global mobility represent new challenges for a global, interconnected community [1]. As diseases affect not only individuals, but also have detrimental effects on whole societies and economies, the prevention and control of infectious diseases is of utmost importance. The ongoing novel coronavirus (SARS-CoV-2) pandemic demonstrates clearly our susceptibility to emerging pathogens.

To respond to the changing environment, the World Health Organization (WHO) released the third edition of the International Health Regulations (IHR) in 2005 in order to strengthen international disease surveillance and control disease outbreaks before they spread [2]. The revised IHR provide an international legal framework for early detection of infectious disease outbreaks by biosurveillance and timely response to them. Early detection is critical to alert health services in a timely manner, and thus to mitigate the impact on morbidity and reduce mortality and economic costs. In this context, public health surveillance is defined by the IHR as “the systematic on-going collection, collation and analysis of data for public health purposes and the timely dissemination of public health information for assessment and public health response as necessary” [2].

1.1.2 Traditional public health surveillance

Traditionally, disease surveillance is carried out by national or supranational public health networks using test results from laboratories [3], [4]. This type of surveillance is also called indicator-based surveillance (IBS). It is based on reporting of individual cases or counts of diseases by sentinel physicians, general practitioners, hospitals, and clinical laboratories. Thus, IBS results in formal and structured data for only a few diseases. The Canadian Chronic Disease Surveillance System (CCDSS) and HIV/AIDS Surveillance System, the American National Tuberculosis Surveillance System (NTSS), and the LaboVIH by Santé Publique France (the French National Institute for Public Health) are all examples of traditional surveillance systems. While these systems are very specific and allow for the estimation of incidence and prevalence, they are resource-heavy, can have a considerable reporting time lag, and lack sensitivity, especially for novel pathogens. For example, the official influenza surveillance data collected by governmental agencies and published by the WHO on the FluNet platform lag behind current flu activity for approximately 2 weeks [5], [6].

1.1.3 Event-based biosurveillance

Due to the limitations of IBS, a multitude of event-based surveillance (EBS) approaches have been developed in recent years, with the goal of near real-time detection of infectious disease outbreaks. A “health event” is defined as any disease outbreak or other occurrence of public health concern [7]. By broadening the focus from specific diseases and counting cases to health events, more timely and complete disease surveillance is possible. EBS has been made possible by two technological advancements in surveillance capacity during the past two decades, namely syndromic surveillance and digital disease surveillance. In order to identify possible outbreaks, syndromic surveillance attempts to detect unusual patterns of health-related events that precede disease confirmation or reporting to official entities [7], [8]. For this purpose, it uses both official and unofficial data such as ICD codes [9], physician billing [10], nurse calls [5], or drug sales [11]. Digital surveillance relies on internet and computer technologies to identify health-related events. For example, Google search queries have been exploited to monitor infectious diseases such as influenza [5], dengue [12], viral gastroenteritis [13], [14], Methicillin-resistant *Staphylococcus aureus* (MRSA) [15], and tuberculosis [16], and Twitter has been used to track influenza [17]–[19] or dengue outbreaks [20]. Syndromic and digital surveillance are not mutually exclusive, but their applications often overlap and complement each other. A key feature of all EBS inputs is that their initial purpose was not biosurveillance. EBS has not been developed to replace IBS, but rather complement it in detecting and identifying health threats [21], a process also called “epidemic intelligence” (EI). However, so far EBS is only used in an informal manner, as the data originating from it have yet to be meaningfully integrated in a formal, quantitative manner.

Approaches to digital disease surveillance vary according to the targeted streams of information, and differ between various biosurveillance systems. Several web-based biosurveillance systems have been developed in recent years by public health organizations and academic institutions [22], [23], such as ProMed Mail [24], GPHIN by Health Canada [25], [26], HealthMap [27], BioCaster [28], MediSys [29], and EIOS by the WHO [30]. The focus in this project will be on data from HealthMap and EIOS due to their availability from public sources and through existing collaborations.

In addition to rapid disease activity detection, internet biosurveillance provides the benefit of greater coverage in regions with fewer medical centers or lower health-seeking behaviors [31]. Moreover, it is cost-efficient because it requires less human curation. Several examples confirm that internet biosurveillance systems are indeed capable of timely disease detection: GPHIN was the first system to detect unusual activity of respiratory illness in the Guangdong Province in China, which later proved to be Severe Acute Respiratory Syndrome (SARS) [25], as well as the 2012 outbreak of Middle East Respiratory Syndrome Coronavirus (MERS-CoV)

[32]. Likewise, ProMed Mail reported the first information on the 2014 Ebola epidemic [33], and monitoring Twitter in Nigeria during the Ebola outbreak in West Africa helped to identify an outbreak three days prior to a news alert and seven days before an official WHO announcement [34].

However, digital disease surveillance systems face their own unique challenges: First, their sources might not be reliable, thus creating false positive signals. Moreover, the sources' signal-to-noise ratio is normally very low, so non-specific information complicates signal detection [35], [36]. Second, the sources may not be sensitive enough to pick up specific disease outbreaks because some diseases are not newsworthy enough. Third, as all systems rely on data from the internet, they are highly dependent on internet coverage and adaption in the countries of operation. Fourth, most systems are heavily language-dependent. For example, HealthMap mainly scans news articles in Arabic, Chinese, English, French, Portuguese, Russian, and Spanish, thus, signals in other languages are missed [37]. Furthermore, since most systems are developed in English, the signal detection sensitivity is greatest in English [38]. Fifth, great media attention for certain diseases or rumors can create false positive signals. Redundancy of information can lead to overestimation of importance, as many news sources report the same events [39]. Sixth, most of the time,

the data is not detailed and reliable enough to provide epidemiological parameters like incidence [40]. Thus, it is necessary to evaluate the EBS systems' performance in general and for specific diseases.

1.2 Influenza surveillance

1.2.1 Influenza as proxy for other diseases

To evaluate how sensitive the systems are in picking up signals on developing outbreaks and how timely they are in detecting these outbreaks, seasonal influenza was used as a proxy for infectious diseases in this project. Influenza was chosen because of its occurrence in multiple countries worldwide, its potential to cause severe epidemics or even pandemics, and its close surveillance and ongoing modeling efforts mostly in rich Western countries [41]. Furthermore, the WHO provides publicly available laboratory-confirmed virological influenza data on a country level on the FluNet platform [42], which can be used as gold standard data for comparisons.

Influenza is tightly monitored because seasonal influenza epidemics are a serious global health threat, causing an estimated annual 3 to 5 million of severe disease cases, and 290,000 to 650,000 respiratory deaths worldwide each year [43]. Moreover, potential emerging influenza strains are a major public health concern, as they could cause influenza pandemics with millions of fatalities.

1.2.2 Virological characteristics

Influenza viruses belong to the family of *Orthomyxoviridae*, which possess a segmented, single-stranded, negative-sense RNA genome. Because the genome is arranged in segments, RNA strands from multiple viral subtypes can be re-assorted, thus generating viruses with novel genetic combinations. This process is called antigenic shift, and can lead to emergence of new viral variants. The other driving factor of emergence of new influenza viruses are frequent mutations, a virological concept also known as antigenic drift [44]. Based on their molecular properties, influenza viruses are subdivided into types *A*, *B*, and *C*. Seasonal epidemics are caused by influenza *A* and *B* viruses, whereas type *C* viruses only cause mild symptoms [44]. Influenza *A* subtypes are further classified by their hemagglutinin (HA) and neuraminidase (NA) surface protein combination (e.g. *H7N9*), and *B* viruses are not further classified, but broken down into lineages (e.g. *Victoria*). So far, pandemics have only been attributed to influenza type *A* viruses [43]. Additionally, influenza *A* infections of poultry, swine, and horses have zoonotic potential [44].

In humans, influenza viruses are rapidly transmitted through infectious droplets or virus-containing aerosols. Once a person is infected, they usually experience a sudden onset of upper respiratory tract illness. Based on the major symptoms, influenza-like illness (ILI) is defined by the CDC as a fever of 100°F (37.8˚C) or greater, cough and/or a sore throat in the absence of a known cause other than influenza [45]. While most people recover from their illness within a week without requiring medical attention, influenza infections can be severe or even fatal in individuals at high risk, such as the elderly, children, pregnant women, and individuals with chronic medical conditions [43].

Influenza infections in temperate regions follow a strong seasonal pattern, where viral activity is centered at one epidemic during the respective winter months. In contrast, in regions close to the equator, influenza infections are endemic. Based on these seasonal patterns, the WHO has defined 18 global influenza transmission regions [46].

1.2.3 Influenza forecasting

IBS data on influenza are predominantly used for monitoring the influenza disease burden, planning and implementation of prevention programs, providing candidate viruses for vaccine production, and resource allocation [42], [47], [48]. Additionally, IBS data are combined with EBS data in models for nowcasting and forecasting influenza activity. Predicted measures typically include season onset, peak week timing, peak intensity, and influenza activity in the next weeks in order to detect unusual or unexpected influenza activity and to prepare the health system for epidemics [41], [49], [50]. Most of the efforts are focused on either North America or Europe [41], [51], [52], such as the CDC flu forecasting challenge [49]. Another example is HealthMap Flu Trends, which publishes real-time estimates of influenza activity in the USA on

a freely available website [53], [54]. HealthMap has also been used as one of seven data sources for short-term forecasting of ILI case counts in South America [55].

1.3 Similar studies

1.3.1 Usage of HealthMap and EIOS

As one of two EBS systems of interest in this work, HealthMap has

been used as an EBS data source for surveillance and forecasting of hantavirus in South America [56], tracking Ebola spread during the West African Ebola epidemic [57] and MERS during the 2012 outbreak [58], and estimating and forecasting Zika virus incidence in South America [59], [60], amongst others. During the ongoing SARS-CoV-2 pandemic, HealthMap provides an interactive map with near real-time updates of geolocated case counts[1], and sounded one of the first alarms of unusual respiratory disease activity in Wuhan on December 30th, 2019, although it was initially dismissed as non-significant [61].

The other EBS system of interest in this work, EIOS, has only been implemented in 2017 and is not publicly available. Thus, only very few publications using EIOS as a data source exist, such as a study by Garten et al, who found that a quarter of infectious disease outbreaks in Africa in 2018 were detected by media monitoring, partly through EIOS [48]. Nevertheless, the EIOS platform is hosted by the WHO and used by public health agencies worldwide [62].

1.3.2 Evaluation of EBS systems

Despite their widespread use, there is not much published evidence about the performance characteristics of EBS systems in terms of disease outbreak detection. Rather, the focus of most of the available literature is on the adequate classification of health-related events from online sources or on the implementation of innovative functionalities [63]. In the first publications about the development of HealthMap, the filtering workflow was described and the accuracy of event classification was assessed, but no validation studies of outbreak detection performance were conducted [64], [65]. Similarly, Dion et al. described the ability of GPHIN to early detect outbreaks of H1N1 and MERS and provide situational awareness (such as information on flight cancellations, border closures, and trade bans) during Ebola and H1N1 epidemics, but did not conduct a formal performance evaluation of GPHIN [32]. This is even more surprising regarding the fact that GPHIN supplied approximately 40% of the WHO's early warning outbreak information in the early 2000s [66]. Similarly, to my knowledge no evaluation or proof of principle of outbreak detection has been published for EIOS.

[1] https://www.healthmap.org/covid-19/

Several authors have criticized the lack of evaluative studies, asking for performance assessment of EBS systems in terms of standardized evaluation criteria against a gold standard [23], and research on the impact of these systems on public health response, epidemic control, and clinical care [67]. In one of these rare studies, Lyon et al. compared the performance of the EpiSPIDER, BioCaster and HealthMap systems in terms of event numbers and distribution, but did not conduct an analysis of outbreak detection functionality [37]. Barboza et al. compared and evaluated six EBS systems operating worldwide, among them HealthMap, in two articles [63], [68]. The first study was focused on the user perspective with a qualitative assessment of metrics such as representativeness, completeness, ease of use, and overall usefulness. Additionally, the authors conducted a quantitative analysis based on detection of H5N1 events in March 2010, in which the detection rate, positive predictive value, sensitivity and timeliness were evaluated [68]. In the second study, the six EBS systems were compared to a gold standard in their ability to detect 23 infectious disease outbreaks, and characteristics associated with detection ability, such as filter languages, types of disease, and regions of occurrence, were identified [63].

In contrast, other web-based sources, such as Google search queries, health-related tweets, or Wikipedia article views, employed by researchers for timely disease detection, are frequently evaluated against traditional surveillance systems regarding their correlation, sensitivity, timeliness, positive predictive value, or forecasting error [19], [67], [69]–[73]. This leads to the paradoxical situation where there is more evidence about experimental data sources and models than about the EBS systems used by official public health entities to guide public health responses and clinical care.

1.4 Objectives

Therefore, as a first objective, this work aims at evaluating the performance of the HealthMap and EIOS systems regarding their ability to detect infectious disease outbreaks. This evaluation is narrow in a sense that it covers only one aspect in which EBS can be used to detect outbreaks. Other approaches like human searching and review of articles are used too, but not covered in this work. Twenty four countries worldwide were chosen to appraise the systems on a global scale and to shift the focus away from rich Western countries, in which most of the disease detection efforts have been concentrated so far. As a proxy for other diseases, influenza will be the agent of interest because it occurs in multiple countries worldwide, albeit with different seasonal patterns, and because laboratory-confirmed influenza counts are available as an accurate gold standard.

As the relative performance of the EBS systems across influenza regions and countries has not been documented, influenza outbreaks in the gold standard and EBS datasets will be identified, and the outbreak detection capabilities of HealthMap and EIOS will be assessed in

terms of sensitivity, specificity, positive predictive value, and timeliness. The metrics will be compared across systems and countries.

As a second objective, factors influencing the detection ability, such as a country's language, wealth, or geographical location, will be identified. The strength of associations of various country-specific characteristics with the evaluation metrics will be assessed in univariable and multivariable regressions.

The results will be of value for the developers of the systems to guide technical improvements. This is especially important for EIOS, since it is still in the development phase. Additionally, the results can guide public health professionals in deciding how the systems can be reliably used for timely detection of influenza epidemics, and potentially epidemics of other infectious diseases.

2 Methods

2.1 Data

2.1.1 Countries

Twenty four countries from 15 influenza transmission zones were chosen to evaluate the performance of event based surveillance (EBS) on a global scale: Argentina, Australia, Brazil, Bulgaria, China, Costa Rica, Ecuador, Egypt, France, Germany, Greece, India, Iran, Mexico, Nigeria, Russia, Saudi Arabia, South Africa, Sweden, Thailand, United Kingdom, United States, Uruguay, and Vietnam. These countries were selected to represent a broad spectrum of geographical locations, languages, and developmental stages.

2.1.2 FluNet: the reference

FluNet is a web-based tool for virological influenza surveillance created by the WHO, and will serve as the reference to evaluate EBS systems. On the FluNet website[2], the WHO provides weekly data on influenza activity by country to the public. These data are gathered from all participating Global Influenza Surveillance and Response System (GISRS) countries, other national influenza reference laboratories which are collaborating with GISRS, and from WHO regional databases [42]. For GISRS, 140 National Influenza Centres around the world collect and test clinical specimens on influenza positivity and submit their results and a sample of these specimens to WHO Collaborating Centres for further characterization [74]. FluNet is used as a tool to explore influenza activity patterns worldwide and guide vaccination programs [42], [47], [75], [76].

FluNet provides publicly available graphs of lab-confirmed influenza cases per country and corresponding csv files since 1997. The csv files include country-specific information such as the WHO region and influenza transmission region, the date and number of received and processed specimen, total number of influenza-positive and negative samples, and a breakdown of these numbers by influenza strain. For the analyses, csv files for 23 countries providing influenza data from January 2013 to December 2019 were downloaded. The only exception was Saudi Arabia, where FluNet data were only available as of January 2017. Since measurements are highly specific, counts from these influenza surveillance data were used as a gold standard against which the EBS system counts were evaluated.

2.1.3 HealthMap

The HealthMap system was developed by researchers at the Boston Children's Hospital, launched in 2006, and provides real-time surveillance on infectious diseases [15]. It collects

[2] https://www.who.int/influenza/gisrs_laboratory/flunet/en/

data from online news aggregators, expert-moderated systems such as ProMed Mail, and validated alerts from official sources such as the WHO. Through automated text processing algorithms in 15 languages, the system filters for disease and location, and publishes the results on a freely available website[3] [12], [20].

Csv files with event data dating from January 2013 to July 2019 were provided by researchers from the HealthMap project group. Any news article which has passed through HealthMap's filtering algorithm and relates to the "human influenza" keyword is referred to as an event. The data files include information about the event's location, the news article headline, a link to the full article, a short article snippet, the news source, and the issue and load date of the article onto the HealthMap platform.

Since every event was identified with a unique HealthMap ID, duplicate events (i.e. articles with exactly the same content, but not different articles about the same event) were removed using this ID information (9012 out of 31796 total events were removed). Additionally, events concerning countries' overseas territories (such as Bermuda, Guadeloupe or Guam) were also discarded, resulting in a total of 22722 unique events in the 24 countries of interest. Since the FluNet gold standard data are only available in weekly intervals, the daily event counts from HealthMap were also aggregated into a weekly format, resulting in a total of 341 weekly data points spanning 6.5 years.

2.1.4 EIOS

Recently, the WHO implemented the Epidemic Intelligence from Open Sources (EIOS) system as a collaboration between multiple public health organizations, acting on a global scale. Its purpose is to integrate data from multiple EBS systems and thus to provide a "unified, all-hazards, One Health approach by using open source information for early detection, verification and assessment of public health risks and threats" [30]. Amongst others, GPHIN, Eurosurveillance, various ministries of health, and big news aggregators are sources for EIOS. Notably, one of the input sources of EIOS is HealthMap, so HealthMap is a proper subset of EIOS. Again, every article that passes through the filtering algorithm and is published on the EIOS platform is referred to as an event. Event de-duplication is performed before the events are uploaded to the website. The EIOS platform is not open to the public, but access was made available through collaborations with the EIOS team. They provided data for every day from November 11, 2017 (the day of EIOS implementation) to December 2019, totaling 109 weeks.

All EIOS events with the following keywords were retrieved and compiled into one file: "Influenza virus not identified" (includes unspecified influenza *A*), "H1N1", "H1N1v", "H1N2",

[3] www.healthmap.org

“H1N2v”, “H2N1”, “H2N2”, “H3N2”, and “H3N2v”. The EIOS file provided data on fetch and import dates of each event, the title, URL and full text of each news article, and information about the news source such as name, country, language, disease category, and mentioned countries. Additionally, an EIOS ID was used as a unique identifier for each event. Since most events mentioned multiple countries, the reports were duplicated so that each row contained only a single country, and then filtered for the 24 selected countries of interest. Additionally, all events without a date stamp were removed from the analysis (320 out of 81133 total events). Like the HealthMap events, EIOS events were aggregated to counts per week per country in order to be able to compare them to the gold standard.

2.1.5 Predictors for regression analysis

Most predictor variables for regression analysis were inherent to the data, such as total counts, geographical locations, and languages. Other predictors were the Human Development Index (HDI) of each country, the total number of internet users (TIU) per country, and the Press Freedom Index (PFI) of each country. HDI rankings from 2018 and the TIU per country in 2017 were obtained from the *United Nations Development Programme* [77]. TIU is expressed as the percentage of a country’s population having access to the internet. The HDI is a composite measure taking into account three dimensions of human development – life expectancy, education, and wealth. It ranges from 0 to 1, with a HDI of less than 0.550 for low human development, 0.550–0.699 for medium human development, 0.7–0.799 for high human development and 0.8 or greater for very high human development. PFI data were downloaded from *Reporters without Borders* [78]. The PFI compiles a number of indicators about freedom of journalism, such as media independence, censorship, and acts of violence against journalists. It ranges from 0 to 100, with higher rankings representing lower press freedom.

2.2 Outbreak detection

2.2.1 Methodological requirements

Outbreak detection had to be performed in the gold standard data as well as in HealthMap and EIOS data because FluNet did not provide a consistent epidemic indicator. Outbreaks were analyzed retrospectively, i.e. the detection method was applied on the whole dataset. The challenge with selecting a method was that for most historical limit approaches, a predefinition of epidemic and non-epidemic phases is needed, which is exactly the desired outcome [79], [80]. In addition to that, the outbreak detection method must not require a long training period, as data are limited, especially for the EIOS system (only 2 years). Moreover, the method should not attempt to model seasonality, as this might mask outbreaks, and will certainly pose problems for application on a global scale, since tropical countries do not show a clear seasonality. Also, influenza seasons are in different months for Northern and Southern hemisphere countries, and some countries have two or more epidemic seasons per year.

2.2.2 Bayesian change point analysis

Therefore, Bayesian change point analysis was the method of choice for outbreak detection. A visual example workflow of the outbreak detection which is described in the following section can be found in Figure 1. Although not initially developed for infectious disease outbreaks [95], [96], change point analysis has been used before to determine start points of influenza epidemics [69], [97]. Essentially, change point methods examine when a change occurs in a series of observations by detecting time points before and after which statistical properties differ. BCP analysis was initially developed by Barry and Hartigan [96] and implemented in R as the package *bcp* by Erdman and Emerson [95], [98]. An excellent theoretical description of the method is provided in those papers.

In BCP analysis, it is assumed that a series of observations is divided into a partition ρ. Each ρ consists of an unknown number of blocks with equal parameter values (Figure 1 top left). Because the implementation in the R package *bcp* assumes a Normal distribution of observation values, the parameters used in this instance are the mean and variance. In order to detect changes, BCP analysis examines the mean of the observations (i.e. weekly number of positive influenza samples for FluNet and weekly number of events for HealthMap and EIOS) before and after a potential change point i. Thus, the time series of observations is divided into b blocks, with the last observation before each block being referred to as a change point. Consequently, a block starts at observation $i+1$ and ends at observation j.

With the Bayesian approach, the objective is to estimate the posterior probability of point i being a change point (p_i), as well as the posterior mean and variance at each i. The prior for the distribution of μ_{ij}(the mean for each block) is chosen as $N(\mu_0, \frac{\sigma_0^2}{j-i})$. This choice of prior requires larger deviations from μ_0 for shorter blocks in order for a change point to be flagged at position i.

Although Barry and Hartigan presented an exact calculation of the parameters, the calculation time for the parameters is *O(n³)*. Therefore, the parameters and positions of change points are estimated with a Markov Chain Monte Carlo (MCMC) approximation, which is only *O(n)*. For each observation i, the following formula is applied:

$$\frac{p_i}{1-p_i} = \frac{P(U_i = 1|X, U_j, j \neq i)}{P(U_i = 0|X, U_j, j \neq i)} = \frac{\left[\int_0^{p_0} p^b (1-p)^{n-b-1} dp\right]\left[\int_0^{w_0} \frac{w^{\frac{b}{2}}}{(W_1 + B_1 w)^{\frac{n-1}{2}}} dw\right]}{\left[\int_0^{p_0} p^{b-1} (1-p)^{n-b} dp\right]\left[\int_0^{w_0} \frac{w^{\frac{b-1}{2}}}{(W_0 + B_o w)^{\frac{n-1}{2}}} dw\right]} \quad (1)$$

where $U_i = 1$ indicates a change point at position $i+1$, and $U_i = 0$ indicates no change point. P is the probability of a change point. p_0 and w_0 are the priors on the probability of a change

point at each position and the signal-to-noise ratio, respectively. b is the number of blocks found if $U_i = 0$. W_1 and W_0 are the within-block sum of squares when $U_i = 1$ and $U_i = 0$, respectively. Likewise, B_1 and B_0 are the between-block sum of squares. $w = \frac{\sigma^2}{\sigma_0^2+\sigma^2}$ where σ_0^2 is the overall variance and σ^2 is the variance of the block observations, so w represents the signal-to-noise ratio. Intuitively, p_i is larger when W_1 small and B_1 is large.

At each step of the Markov chain, a value for U_i is drawn from the conditional distribution of U_i, given the data and the current segmentation of observations into blocks. Based on these values, p_i is calculated with ((1). After every MCMC iteration, p_i and the posterior mean of each block are updated. The posterior mean for every block is calculated with the formula $\hat{\mu}_{ij} = (1-w)\bar{X}_{ij} + w\mu_0$, where $w = \frac{\sigma^2}{\sigma_0^2+\sigma^2}$ and $\bar{X}_{ij}$ is the mean of the observations in the block b_{ij}.

All analyses were conducted using the R package *bcp*, version 4.0.3 [98]. Assumptions of *bcp* are that observations are distributed independently $N(\mu_i, \sigma^2)$ in different blocks, given the parameters, and that p_i is independent at each observation, conditional on the partition. In contrast to the assumptions, influenza case counts and EBS event counts followed an over-dispersed Poisson rather than a Normal distribution. However, counts during epidemic and non-epidemic periods obviously have different means, so BCP analysis is able to clearly separate them from each other. Nevertheless, BCP usually detected several changes in mean during epidemic periods, so p_i is high for many points during rising and falling epidemic curves.

The posterior means $\hat{\mu}_{ij}$ and the posterior probabilities of a change point p_i were estimated with 600 MCMC iterations, where the first 100 MCMC iterations were discarded as burn-in (Figure 1 top right). The tuning parameters of the priors p_0 (change point probability) and w_0 (signal-to-noise ratio) were kept at their default value of 0.2. These values have been found to work well in the past [96], and neither HealthMap nor EIOS data were sensitive to changing priors (see figures **Figure 9Figure 10** in the appendix).

2.2.1 Determination of start and end points of epidemics

As multiple change points were flagged during outbreaks (Figure 1 bottom left), criteria were established to determine the start and end points of influenza outbreaks. The criteria were as follows:

- Start points: rising count curve, no outbreak start flagged during previous 15 weeks, transition from p < 0.5 to p > 0.5 (i.e. first change point after a non-epidemic period).
- End points: falling count curve, no outbreak end flagged during preceding 15 weeks, transition from p > 0.5 to p < 0.5 (i.e. last change point of an epidemic period)

0.5 is the intuitive and widely used posterior probability threshold for determining change points. Nevertheless, receiver operating characteristic (ROC) curves of a sequence of thresholds were plotted to check for the best threshold. For low-count countries, thresholds did not matter at all, while for higher count countries, 0.5 was a good compromise between all countries (see **Table 7** in appendix). For reasons of simplicity and generalizability, cutoffs at the optimal thresholds for each country were not established.

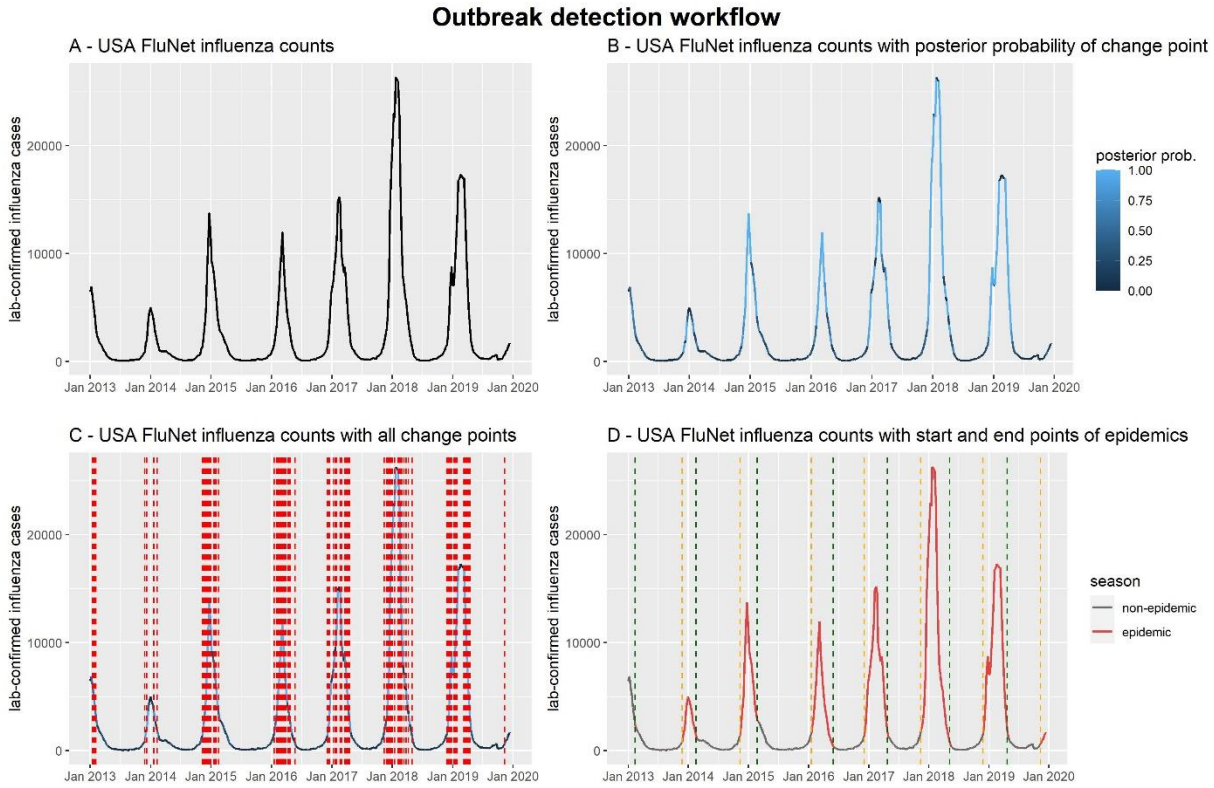

Figure 1: Workflow of outbreak detection with BCP exemplified for one country. *(A) The time series of weekly aggregated influenza cases for the USA is plotted. This is the raw data. (B) The observations are colored according to the posterior probability of a change point at each observation. This step is after running the BCP analysis. (C) The position of all change points (posterior probability > 0.5) is visualized with vertical dashed lines. (D) After applying criteria for start and end points of epidemics, the time series is divided into epidemic (red) and non-epidemic (grey) seasons. Start points of epidemics are plotted as orange vertical lines, end points as green vertical lines.*

In order to disregard local spikes or drops in the count data, rising and falling curves were determined by smoothing the time series with Loess smoothing. The window size used for smoothing of FluNet and HealthMap data was chosen as 10%, i.e. 10% of the whole dataset were taken into account, which corresponded to 36 and 34 weeks, respectively. Since the EIOS dataset comprised only 109 weeks, a window size of 15% (16 weeks) was chosen. These time intervals showed a good compromise between smoothing over local spikes or drops in the counts and overall variation in counts.

The 15 week period of no start points before a start point was chosen so that the algorithm would not flag epidemic starts during an ongoing outbreak. On average, influenza epidemics

were found to be 3.9 months long, irrespective of climatic regions, with a range of 3-5 months [75]. Regarding the length of the epidemic period and assuming that there is some amount of non-epidemic time between two outbreaks, the period between to 'ends' was chosen to be at least 15 weeks as well (Figure 1 bottom right).

When detecting spikes, the algorithm usually flagged an epidemic start point, but often no end point. In order to allow detection of spikes as single outbreak weeks, an 'end' was inserted after every 'start' that was not followed by an 'end' within a period of 30 weeks (which is safely over the 5 months maximum epidemic period found in by Azziz-Baumgartner et al. in [75]).

2.3 Performance evaluation metrics

The EBS system performance was evaluated regarding sensitivity, specificity, positive predictive value, and timeliness. Sensitivity was measured in three ways: Sensitivity per outbreak to evaluate overall detection of outbreaks (equation (2), sensitivity per week to be able to calculate composite metrics (equation (3), and timely sensitivity to combine timeliness and sensitivity for detection of outbreaks within two weeks of the start of an epidemic (equation (4). This last metric is the most relevant because the timely detection of outbreaks is what the EBS systems were designed for. A window of two weeks before and after an outbreak was detected in the gold standard data was chosen so that an alarm would be raised before or at the time a traditional surveillance system would detect an anomaly. In all equations, "outbreak" refers to a detected outbreak in the gold standard data and "alarm" refers to a detected outbreak in the EBS system data.

$$sensitivity\ per\ outbreak = \frac{n\big(alarm\ at\ any\ time\ during\ outbreak\big)}{n(outbreak)} \quad (2)$$

$$sensitivity\ per\ week = \frac{n\big(alarm = 1,\ outbreak = 1\big)}{n(outbreak = 1)} \quad (3)$$

$$timely\ sensitivity = \frac{n\big(alarm\ at\ \pm 2\ weeks\ of\ outbreak\ onset\big)}{n(outbreak)} \quad (4)$$

Timeliness was calculated as the mean of prevented fraction of outbreaks

to circumvent the problem of non-detected outbreaks. The prevented fraction is the proportion of time of an outbreak saved by detection by the EBS system relative to the onset of the outbreak [99]. If an outbreak is detected, it is calculated as:

$$timeliness\ (prevented\ fraction) = 1 - \frac{t_{alarm} - t_{onset}}{outbreak\ duration} \quad (5)$$

where t_{onset} is the onset time of the outbreak, t_{alarm} is the time of detection by the EBS system and $outbreak\ duration$ is the number of weeks for which an outbreak continues. Thus, timeliness will be 1 if an alarm is raised by the EBS system in the onset week and 0 if an alarm is raised at the end of an outbreak or if the outbreak is not detected by the EBS system at all. As all datasets are comprised of multiple outbreaks, the arithmetic mean of all prevented fractions for each system and country were reported.

Specificity was calculated per week (equation (6), as was the positive predictive value of alarms (equation (7). Additionally, sensitivity and specificity per week were combined into an accuracy metric, which is the sum of all correctly assigned weeks over the total number of weeks (equation (8). Moreover, the F1 score was calculated as the harmonic mean of sensitivity per week (i.e. precision) and positive predictive value (i.e. recall) (equation (9).

$$specificity\ per\ week = \frac{n(alarm = 0,\ outbreak = 0)}{n(outbreak = 0)} \quad (6)$$

$$positive\ predictive\ value\ (PPV) = \frac{n(alarm = 1,\ outbreak = 1)}{n(alarm = 1)} \quad (7)$$

$$accuracy = \frac{n(alarm = 1,\ outbreak = 1) + n(alarm = 0,\ outbreak = 0)}{n(weeks)} \quad (8)$$

$$F1 = 2 \times \frac{precision\ (PPV) \times recall\ (sensitivity\ per\ week)}{precision\ (PPV) + recall\ (sensitivity\ per\ week)} \quad (9)$$

95% confidence intervals for the evaluation metrics were calculated as exact binomial confidence intervals, with the exception of timeliness. Since timeliness was calculated as the mean of the prevented fractions per outbreak, a binomial procedure was not possible. Thus, similarly to the procedure of calculating confidence intervals for odds ratios, the point estimates were logit-transformed, and the standard deviation was determined as $\sqrt{\frac{prev.fraction \times (1-prev.fraction)}{n(outbreaks)}}$. The 95% confidence intervals were then calculated on the logit scale and back-transformed to the linear scale, thus being bounded between 0 and 1.

2.4 Regressions

Country-specific variables examined as predictors in regressions to test their influence on EBS system performance were: total counts over the data collection period, maximum counts per week, global region (temperate Northern hemisphere, temperate Southern hemisphere or tropical), language (English yes/no), latitude, longitude, human development index, press freedom index, total numbers of internet users and HealthMap filter language (yes/no). No variable for EIOS filter language was set up because EIOS filters in the language of each test

country. The logarithm of total and maximum counts was used in regression analysis to ensure linearity between outcome and predictor. The absolute value of latitude was used as a proxy for a country's climate. All regressions were fit in the software R (version 3.6.3) [100], using the packages *base*, *MASS* (version 7.3-51.6) [101], and *glmnet* (version 4.0) [102]. Additional diagnostic plots were created with *car* package version 3.0-7 [103].

First, to determine the influence of certain country-specific factors on the EBS system performance, univariable regressions were performed with each evaluation metric as outcome and each covariate as a predictor. All assumptions of linear regressions were checked with diagnostic plots. These are: histograms of the residuals to check the Normality assumption, plotting of fitted values against residuals for checking for a linear relationship and homoscedasticity, QQ-plots of the residuals to check for Normality, and leverage plots for outliers. The assumptions of normally distributed errors and linearity between outcome and predictor were not met for timely sensitivity, as zeros were overrepresented in the values. Thus, logistic regressions were performed to see which factors influenced if timely sensitivity was 0 or above 0.

Next, predictors with a p-value below 0.2 in univariable regressions were included in multivariable linear regressions for each evaluation metric, and the variance inflation factors were checked with the R package *car* (version 3.0-7) [103]. Highly correlated variables were removed from the models until the variance inflation factors were below 4, which is a conservative threshold [104]. Then, for each outcome, influential variables were selected into a final model by a forward selection process, which was based on the models' Akaike information criterion (AIC).

To see if the variable selection process by AIC was valid, the selection process was repeated using least absolute shrinkage and selection operator (LASSO) regression. However, only the coefficients from the AIC models were reported, since the LASSO coefficients are biased due to shrinkage. LASSO regression was carried out with the *glmnet* package. The parameter for the amount of the coefficient shrinkage λ was optimized to the value that minimized the cross-validation prediction error rate. The subset of variables whose coefficients were not shrunk to 0 was then compared to the set of variables obtained through selection with the AIC criterion.

3 Results

3.1 FluNet Data

Confirmed virological influenza case counts from over 7 years were collected from FluNet, a traditional influenza surveillance platform. In most countries, the data were abundant and of good quality, although the absolute number of counts varied greatly across countries, reflecting varying testing capacities between countries. While countries fully situated in temperate regions showed one distinct epidemic per year in their respective winter months, some countries spanning multiple climate zones such as India and China showed two epidemic peaks in some years (Figure 2). Countries situated in tropical regions such as Costa Rica and Ecuador displayed more irregular patterns with missing one season or multiple outbreaks per season. In three tropical countries with no clear seasonality (Nigeria, Thailand, and Vietnam), FluNet data were of low quality. The total number of tested individuals and thus positive influenza cases was very low and the background noise was so high that almost no outbreaks were discernible.

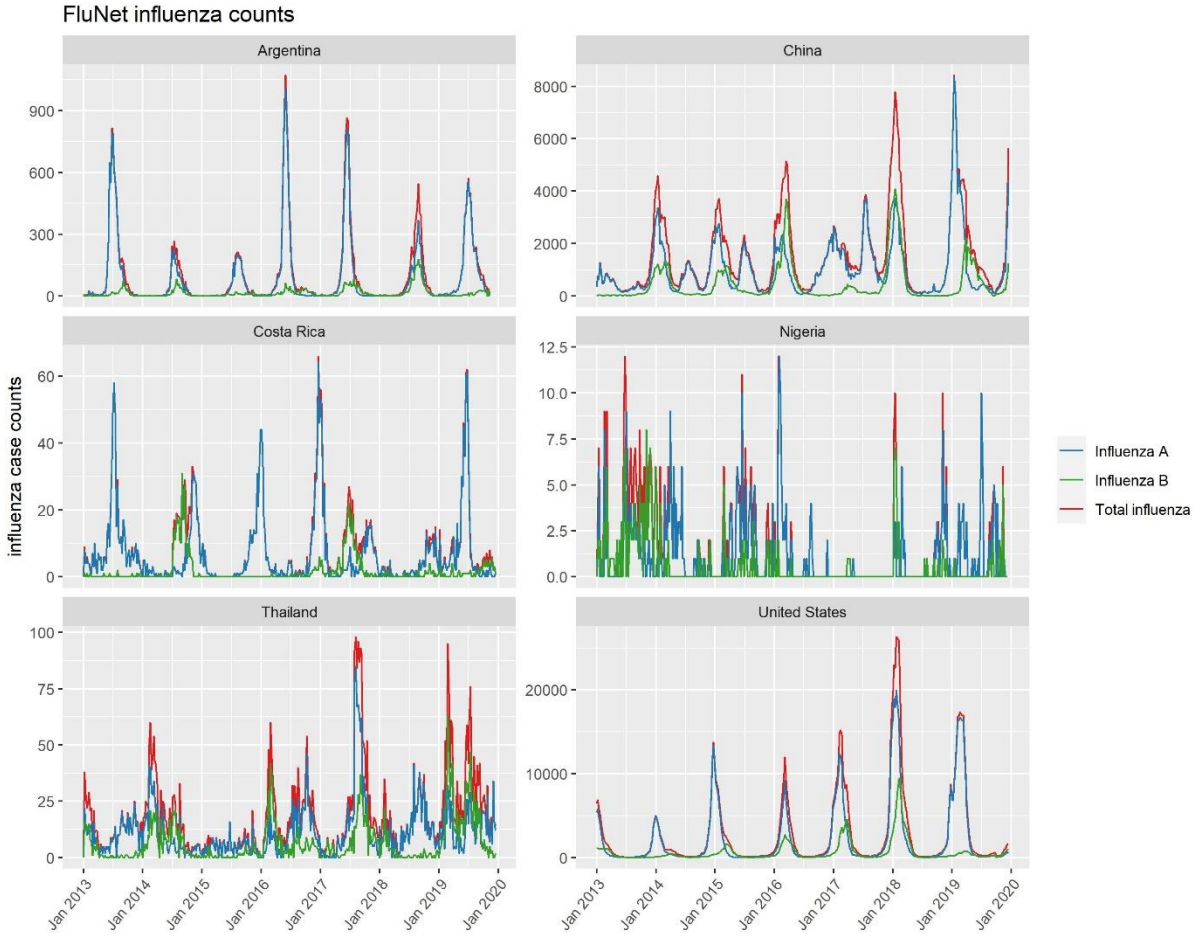

Figure 2: Positive case counts for influenza types A, B and total from the FluNet platform plotted as a function of time for 7 years. *6 representative countries from different influenza transmission zones were chosen to visualize a broad spectrum of seasonal influenza patterns, testing capacities, and data quality.*

Plotting FluNet counts over time revealed an interesting interplay between influenza subtypes *A* and *B*. In general, influenza *B* cases peaked at a slightly different time than influenza *A* and case counts were lower, except in the European countries in the season 2017/2018. In some years, influenza *B* cases preceded or lagged behind the peak of influenza *A* cases greatly, leading to bimodal peaks of total influenza counts, with one peak for influenza *A* and one peak for influenza *B*. Another insight from the FluNet data is that influenza cases have a high year-to-year variability: they do not always peak at the same time of the year, nor do the peaks always have the same height, indicating that influenza epidemics are worse in some years than in others.

Despite data limitations in three countries, total influenza counts were used as gold standard for the following analyses of event-based surveillance (EBS) system performance. Influenza *A* and *B* could not be regarded separately because news articles rarely distinguish between influenza subtypes.

3.2 EBS total count data

Influenza events were gathered from HealthMap for 6.5 years and from EIOS for 2 years. The average annual number of HealthMap events was lower than the average annual number of events produced by the EIOS system in every country (Figure 3). This difference can be explained by the fact that EIOS receives a great deal more input than HealthMap by not only scraping news aggregators, but also aggregating multiple EBS systems (HealthMap and GPHIN amongst others).

In both systems, the USA was an outlier with the highest number of events by far. Twelve countries had only very sparse event data in HealthMap (Bulgaria, Costa Rica, Ecuador, Germany, Greece, Iran, Nigeria, Saudi Arabia, South Africa, Sweden, Thailand, and Uruguay). Surprisingly, these low count countries were not all developing countries with low internet usage and poor health systems, but also included rich, Western countries. However, this observation can be explained by the fact that HealthMap does not filter for news articles in the official languages of these countries. In contrast, EIOS filters for all the official languages of the 24 countries evaluated, thus filtering language cannot be a factor affecting EIOS count numbers. Nevertheless, EIOS also shows large differences in total counts between countries, with low count countries mostly situated in tropical regions.

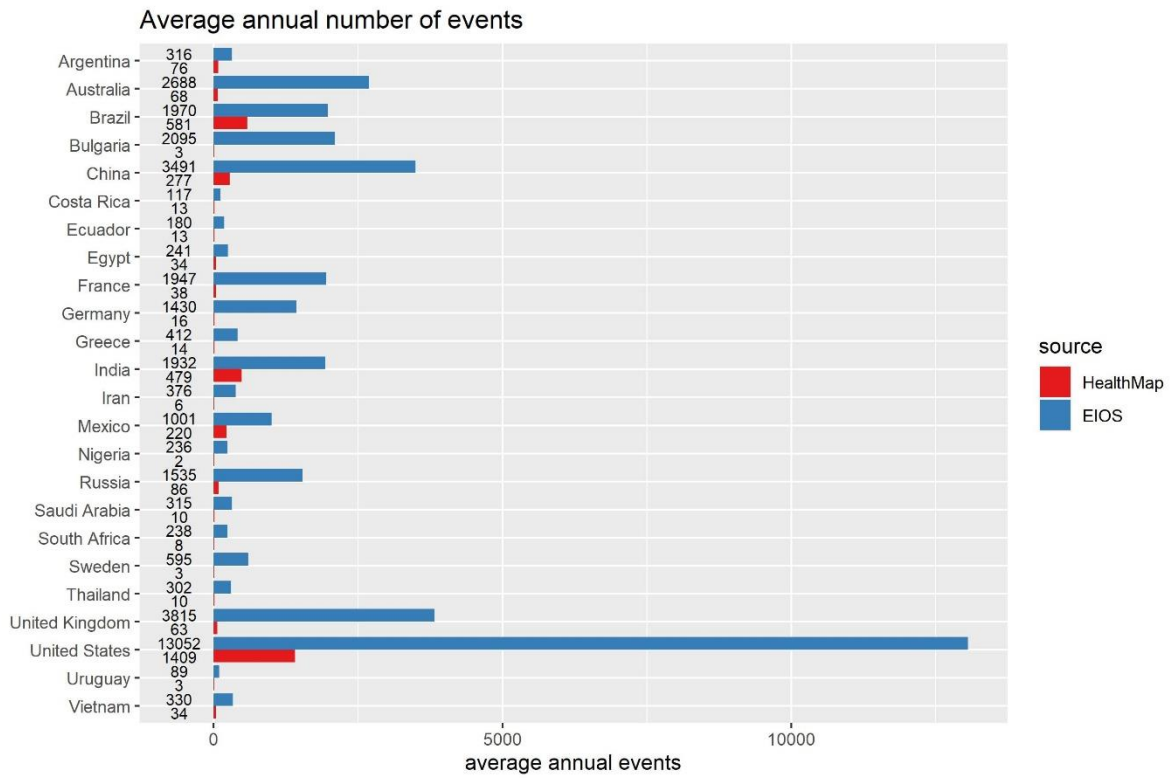

Figure 3: Average annual counts per country of EIOS and HealthMap systems over 2 and 6.5 years, respectively. *Every news article picked up by the systems is counted as an event.*

3.3 Visual correlation between EBS systems and gold standard over time

The EBS systems reflect the seasonal influenza outbreaks detected in FluNet counts to some extent. Figure 4 exemplifies the visual correlation between FluNet, HealthMap and EIOS counts for a set of countries with very different numbers of events.

Countries with very low event counts in HealthMap show only scattered spikes of a few events. However, most of these spikes occur when confirmed influenza cases are high, hinting at some degree of correlation (such as Bulgaria and Germany), but they can also occur seemingly at random (such as Nigeria). On the other hand, HealthMap events in countries with higher event counts generally coincide with influenza epidemics in FluNet data, as in Argentina, India and the USA.

In contrast, EIOS events seem to be less synchronized with FluNet counts and have a lower signal-to-noise ratio. Additionally, event count patterns are less clearly visible, as EIOS provided event data for only 2 years. Exceptions are only countries with a very high number of EIOS events, such as Bulgaria and the USA, which show a distinct seasonal pattern.

There are also noteworthy differences in number of events for the two systems between different countries. For instance, in Bulgaria, HealthMap collected only very few events, whereas EIOS found many events that were well-correlated with the gold standard. Looking only at the visual correlations, it is already clear that the two EBS systems operate with very

different characteristics, which lead to different distributions of event counts: EIOS data were more abundant, but more variable and were less synchronized with lab-confirmed influenza cases. In contrast, HealthMap generally did not produce any events at all during non-epidemic periods and produced spikes or curves of only moderate height during epidemic periods.

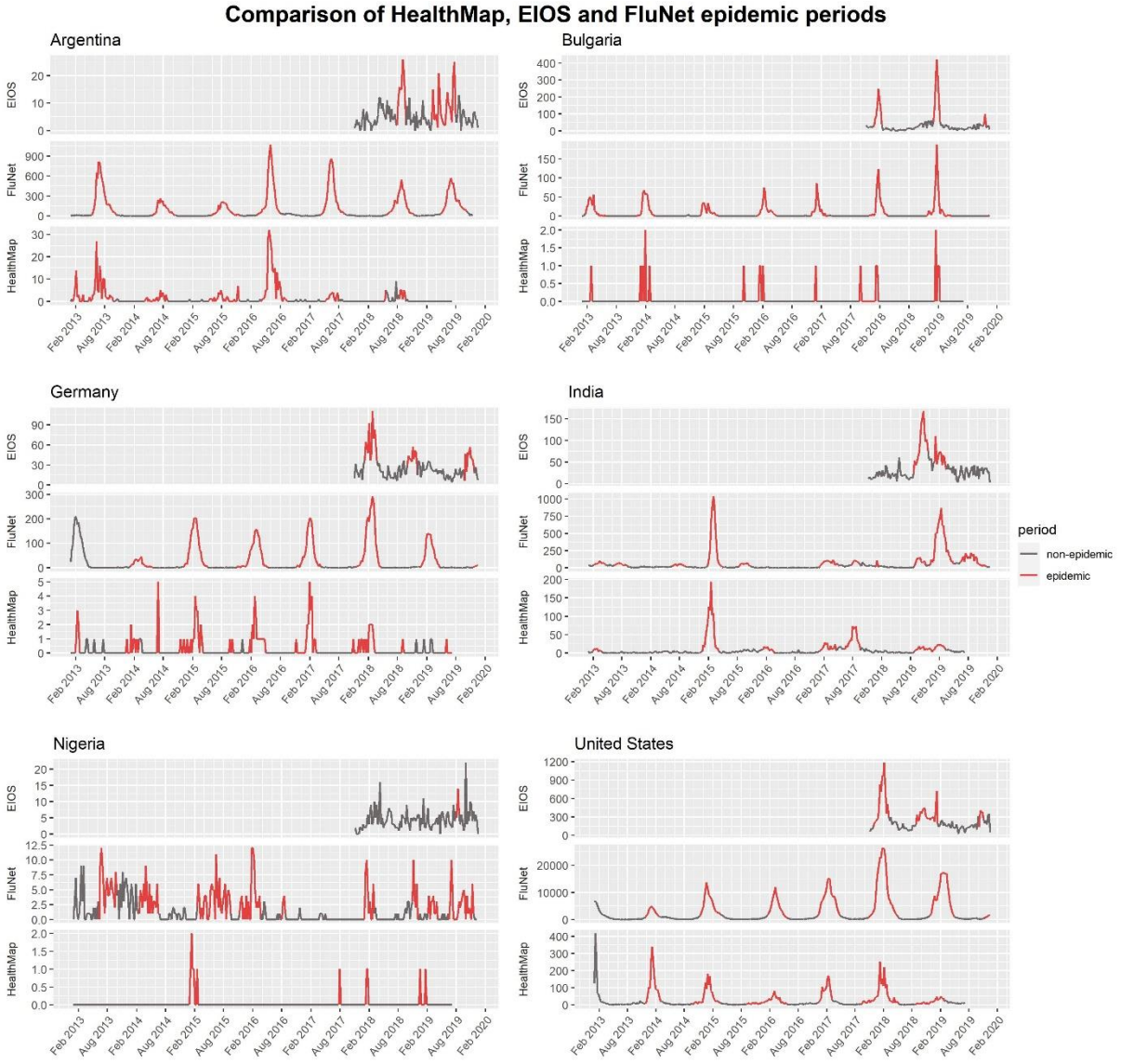

Figure 4: Time series correlations for EIOS, FluNet and HealthMap. *Weekly events relating to influenza from every system are plotted over the time that data are available. EIOS events are on top, confirmed influenza counts from FluNet in the middle, and HealthMap events at the bottom. Epidemic periods found with Bayesian change point analysis are highlighted in red, non-epidemic periods are shown in grey.*

3.4 Outbreak detection

To go beyond the visual correlation, start and end points of influenza epidemics were determined retrospectively by Bayesian change point (BCP) analysis in all three datasets. Based on the detected start and end points, the time series were divided into ‘epidemic’ and ‘non-epidemic’ periods.

3.4.1 FluNet

Regular seasonal epidemics were discovered in all countries except in Nigeria, Thailand and Vietnam, for which FluNet provided only sparse data and which, due to their tropical climate, inherently show no clear seasonality. In line with the observations made by Azziz-Baumgartner [75] and Newman [47], Brazil, China, Costa Rica, Ecuador, Egypt, India, and Mexico experienced a second epidemic in some years, and Nigeria, Thailand and Vietnam showed year-round activity. Most countries had 8 or 9 outbreaks during the whole study period. The maximum number of outbreaks detected was 11 (in India and Nigeria), and the minimum number was 4 in Saudi Arabia, since data were only available as of January 2017. The peak height for outbreaks varied significantly between countries, but also between outbreaks within countries. The countries with the highest number of confirmed cases per outbreak were the USA, China, and France, while the countries with the lowest peaks were Nigeria, Uruguay, and Vietnam.

The epidemic periods of the FluNet counts displayed in figure 3 confirmed the visual detection of outbreaks in the timeline very well. BCP analysis detected a change early in rising curves and at the end of dropping curves. For Nigeria, Thailand and Vietnam, the epidemic periods detected by BCP are less well-defined due to the high signal-to-noise ratio of influenza counts in these countries (Figure 5). Therefore, epidemic indicators cannot be fully trusted and performance metrics which are calculated later may be questionable for these countries.

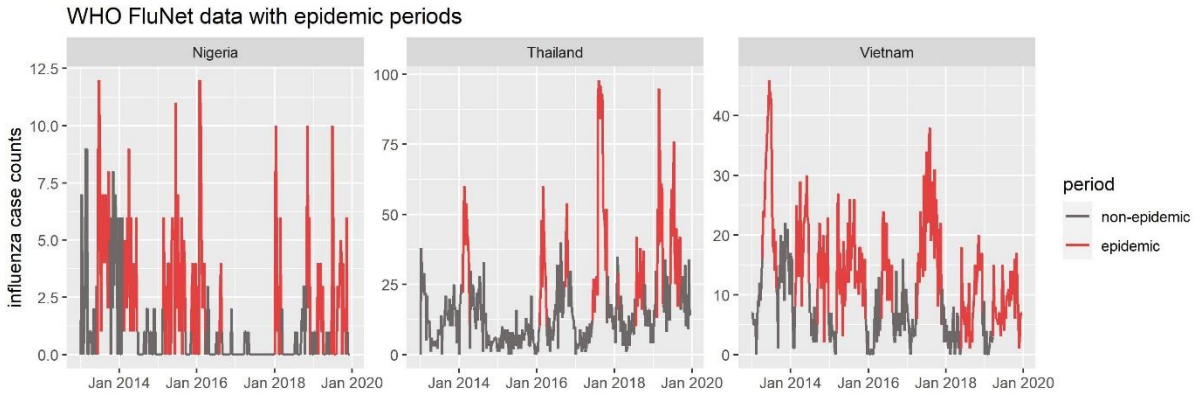

Figure 5: FluNet data with epidemic period indicator for the three countries with low data quality. *Epidemic periods, as defined by Bayesian Change Point approach, are colored in red, non-epidemic periods in grey.*

3.4.2 EBS systems

Twelve countries had very low HealthMap event counts and thus only showed spikes of a few events at a time. BCP analysis detected all of these spikes as outbreaks (see data for Bulgaria, Germany, and Nigeria in Figure 4). Detecting these spikes as outbreaks increased sensitivity compared to ignoring them (see section 3.5.3). In line with observations from FluNet, outbreaks

detected by the BCP algorithm in HealthMap data corresponded well with visual outbreak detection.

Due to the low signal-to-noise ratio and the short data collection period of EIOS, a clear baseline could not be discriminated from an epidemic phase for this system. Therefore, it seems that BCP analysis did not work not as well as in HealthMap. Most outbreaks are still recognized as they would with the naked eye, but sometimes the algorithm did not detect spikes (see Nigeria in Figure 4) or flagged outbreaks at unexpected time points (see Argentina in Figure 4). Moreover, EIOS counts peaked at times when gold standard counts were still low, for example in Germany in the fall of 2018 and 2019 or in the USA in the summer/fall of 2018.

3.5 Evaluation of outbreak detection performance

Datasets were first divided into epidemic and non-epidemic periods. Influenza outbreak detection performance was evaluated separately for each EBS system because both have their unique characteristics and biases. To obtain a complete picture of the performance, three different metrics of sensitivity were evaluated along with specificity, positive predictive value, and timeliness of detection. Additionally, accuracy and F1 scores were calculated as two composite measures of performance.

In general, system performance varied widely across countries, and there was seldom a discernible concordance between EIOS and HealthMap. Of note, EIOS metrics were less precise than HealthMap metrics because the former are calculated from only two years of data, whereas HealthMap calculations are based on 6.5 years of data. Hence, the 95% confidence intervals of all EIOS metrics are wider than their respective HealthMap counterparts (**Table 8** in appendix).

3.5.1 Performance measured in simple metrics

Sensitivity per outbreak was over 50% for most countries in HealthMap, except for Egypt and Nigeria, and over 75% for 13/24 countries (Figure 6). That is to say, in 13/24 countries HealthMap detected ¾ of all outbreaks. In comparison, EIOS detected ¾ of all outbreaks in only 9/24 countries and did not detect a single outbreak in Costa Rica. Sensitivity per outbreak was the evaluation metric which had the highest number of countries scoring 100%, that is, 6 in HealthMap and 9 in EIOS. Remarkably, EIOS thus had very high inter-country variations in this metric.

Sensitivity per week was lower in both systems for every country, except obviously for Costa Rica in EIOS, which was again 0%. For example, while HealthMap achieved a sensitivity per outbreak of 86% in Argentina, the country's sensitivity per week was only 63%. In EIOS, the discrepancy for Argentina was even higher, with a sensitivity per outbreak of 100% and a

sensitivity per week of 39.5%. Since the main reason for the calculation of this metric was the usage in composite measures, the scores of individual countries will not be discussed.

Neither of the EBS systems achieved a good score for timely sensitivity on average, which was the main outcome of interest. HealthMap detected 0% of all outbreaks within 2 weeks of outbreak onset in 13/24 countries, and EIOS failed to detect outbreaks in a timely manner in 19/24 countries (Figure 6). The countries in which HealthMap detected the most outbreaks on time were Bulgaria, the UK and the USA with a timely sensitivity of 28.6%. This corresponds to 2 outbreaks out of 7 which were detected within 2 weeks of outbreak onset. For EIOS, the timely sensitivity of France and Vietnam was 50%, corresponding to 1 outbreak out of 2 detected within two weeks, and 1 out of 3 outbreaks was detected in Brazil, Sweden, and the UK.

The specificity of both systems did not vary substantially between countries and was generally very high, with 22 countries having a specificity larger than 75% in HealthMap and 17 in EIOS. Iran in EIOS was the only country which had a specificity of 100%, meaning that all weeks in which an alarm was raised were classified as outbreak weeks in the gold standard, too.

Calculation of the positive predictive value produced more heterogeneous results. In HealthMap, 4 countries had a PPV below 50%, 11 between 50% and 75%, and 9 above 75%. For example, this means that the probability of a week being classified as an outbreak in HealthMap also being classified as an outbreak week in the gold standard data was less than 50% for France, Nigeria, Saudi Arabia, and Thailand. In EIOS, the PPV ranged from 0% in Costa Rica to 100% in Iran, with 8 countries below 50%, 7 between 50% and 75%, and 8 above 75%.

Timeliness was calculated as the mean of prevented fractions of outbreaks to circumvent the problem of non-detected outbreaks. For both systems, the mean prevented fraction was rarely over 75%, and especially low in EIOS with 15 countries below 50%. This means that the systems usually detected outbreaks late after their onset, which corresponds well to the low scores of timely sensitivity. Countries in which the prevented fractions are high are Argentina, the United States, and Vietnam for HealthMap, and Argentina, China, France, Sweden, and the UK for EIOS.

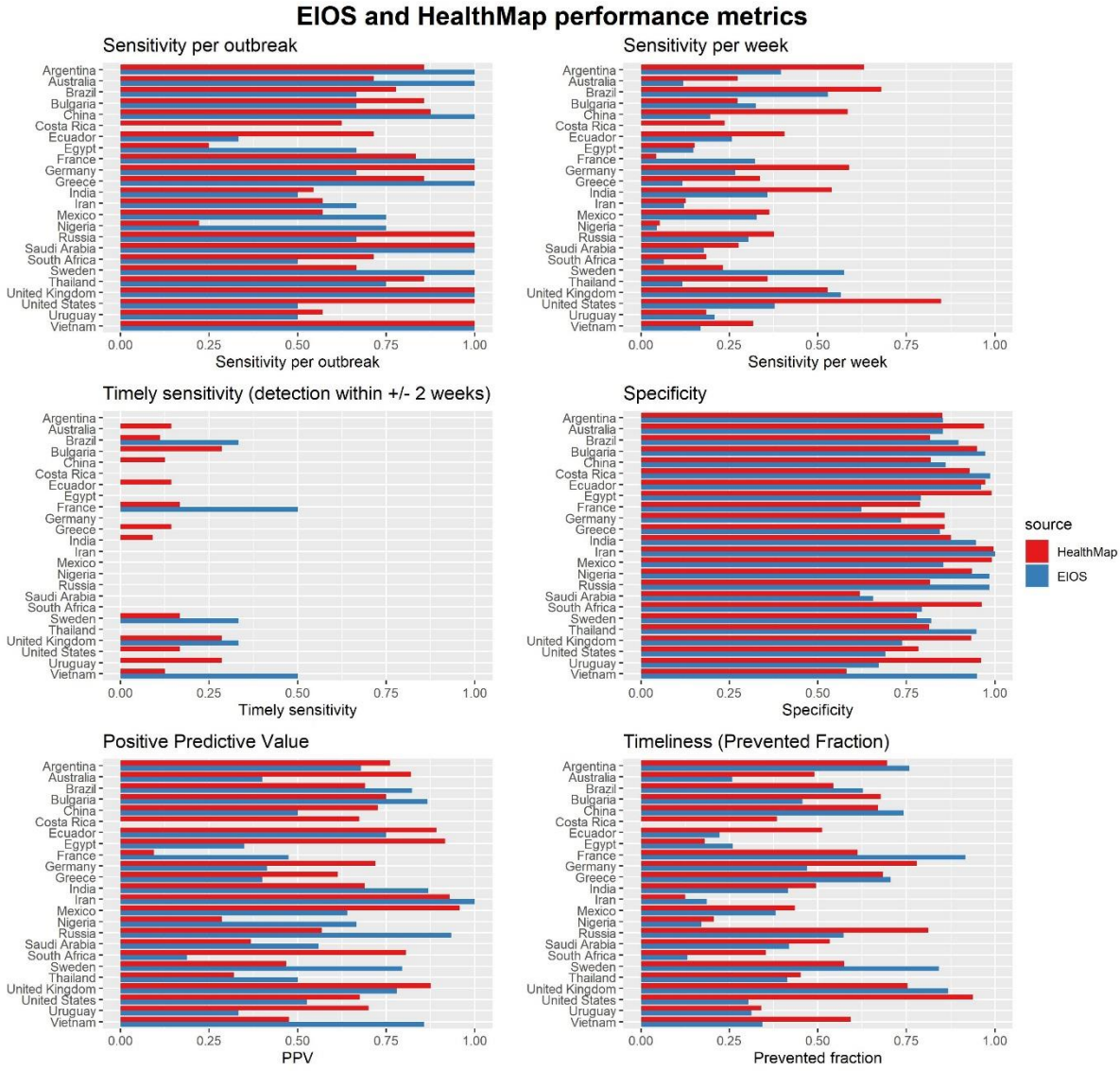

Figure 6: EIOS and HealthMap performance metrics. *All metrics were calculated with FluNet data as reference. Sensitivity per outbreak is the sum of all detected outbreaks (alarm during the outbreak) over the total number of outbreaks. Sensitivity per week is the sum of correctly classified epidemic weeks over the total number of epidemic weeks. Timely sensitivity is the sum of all outbreaks detected within 2 weeks of the onset date in FluNet divided by the total number of outbreaks. Specificity is the sum of all correctly classified non-epidemic weeks over the total number of non-epidemic weeks. Positive predictive value is the number of weeks correctly classified as epidemic divided by all epidemic weeks. Timeliness or prevented fraction is defined as the difference between the outbreak onset and the time of detection by the system, divided by the length of the whole outbreak. If an outbreak is not detected by the system, the prevented fraction is set to zero.*

3.5.2 Performance measured in composite metrics

In addition to these simple metrics, two composite measures were calculated: accuracy is the number of correctly classified weeks over the total number of weeks, so it combines sensitivity per week and specificity. For HealthMap, accuracy was 75% or higher in 7 countries and 50% or higher for all countries except Saudi Arabia and Vietnam (Figure 7). Countries for which HealthMap was over 75% accurate were the United States, Ecuador, Brazil, Argentina, Mexico, India, and Germany. In EIOS, accuracy was over 70% for Brazil, Russia, Bulgaria, and Ecuador, and the countries with the lowest accuracy were Saudi Arabia and Vietnam.

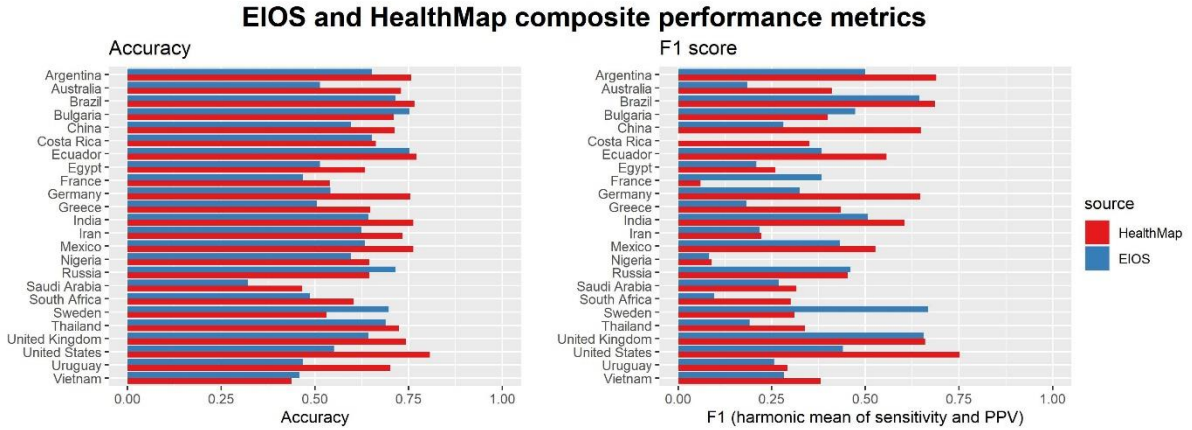

Figure 7: EIOS and HealthMap composite performance metrics. *All metrics were calculated with FluNet data as reference. Accuracy is the sum of all correctly classified weeks over the total number of evaluation weeks. The F1 score is the harmonic mean of sensitivity per week (recall) and positive predictive value (precision).*

The F1 score was calculated as the harmonic mean of sensitivity and PPV, and was higher most of the time in HealthMap than in EIOS. In HealthMap, the USA and Argentina had the highest F1 score with 75% and 69%, respectively, and Nigeria and France had the lowest scores with 9% and 6%, respectively. Sweden, the UK, and Brazil were the only countries which scored over 60% in EIOS. The countries with the lowest scores in EIOS were South Africa, Nigeria, and Costa Rica with F1 scores of 9.7%, 8.3% and 0%, respectively. Costa Rica had a F1 score of 0%, since its sensitivity was 0%. In general, these results suggest that HealthMap performed consistently better than EIOS, with higher average values in all performance metrics (Table 1: Summary of EIOS and HealthMap performance).

	Sensitivity per outbreak	**Timely sensitivity**	**Sensitivity per week**	**Positive predictive value**	**Specificity**	**Prevented fraction**	**Accuracy**	**F1 score**
HealthMap > EIOS	11	8	19	13	11	15	20	20
HealthMap = EIOS	3	11	0	0	0	0	0	0
EIOS > HealthMap	10	5	5	11	13	9	4	4
HealthMap average	0.753	0.093	0.359	0.657	0.869	0.535	0.677	0.433
EIOS average	0.733	0.083	0.254	0.596	0.850	0.449	0.591	0.338

Table 1: Summary of EIOS and HealthMap performance.

3.5.3 Comparison system performance with a count limit

In order to check if classifying single spikes of very few events as outbreaks increases system performance, an analysis was conducted in which an outbreak was only flagged with a minimum of 5 weekly counts. This analysis was only carried out for HealthMap data, as EIOS did not show such spikes. Introducing an event count limit for outbreak detection reduced the average specificity and PPV, but increased sensitivity per week and timeliness (

Figure **8**). Timely sensitivity and sensitivity per outbreak were unchanged. On average, the limit also decreased the F1 score significantly and had no effect at all on the accuracy. Therefore, it can be concluded that a count limit is not a useful method to improve detection of influenza outbreaks.

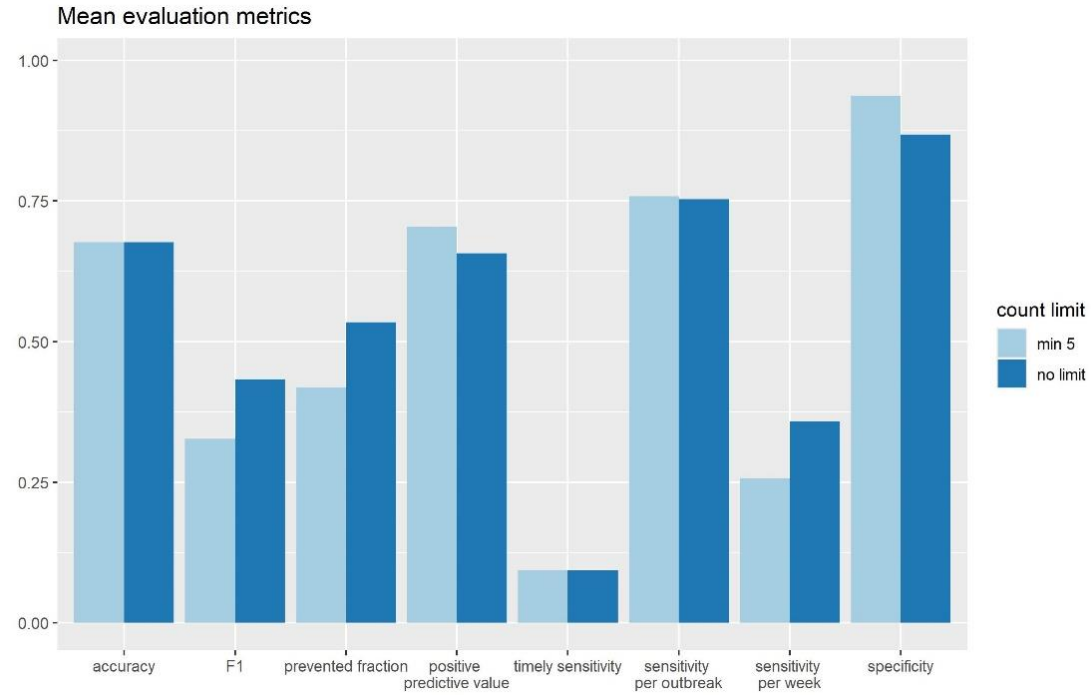

***Figure 8: Comparison of mean evaluation metrics for HealthMap with and without an outbreak detection limit of at least 5 weekly counts**. All metrics were calculated with FluNet data as reference.*

3.6 Detection of country factors influencing system performance

3.6.1 Correlations between outcomes and predictors

Before country-specific factors that influence HealthMap and EIOS performance could be identified, scatter plots of all predictors and all evaluation metrics were examined. Not surprisingly, most of the evaluation metrics were correlated with each other (**Figure 11Figure 12** in appendix). All sensitivity metrics were moderately positively correlated with each other, as were sensitivity and prevented fraction. Moreover, specificity was positively correlated with PPV and negatively correlated with sensitivity per outbreak, sensitivity per week, and prevented fraction. However, specificity was not correlated with timely sensitivity, as the latter was comprised of too many zeros.

Some predictors were strongly positively correlated with each other, such as total and maximum counts of each EBS system, or HDI and total internet users (**Figure 13** in appendix). Moderate correlations existed between HDI and latitude, total internet users and latitude, HDI and PFI, and latitude and EIOS counts. Interestingly, HDI was moderately correlated with EIOS event counts but not with HealthMap event counts. Not surprisingly, HealthMap and EIOS

counts were correlated, too. Because of the high degree of correlation, some variables were removed from the multivariable regression models. Variance inflation factors were lowest when total counts (as categorical variable), maximum counts, global region (as categorical variable), and total number of internet users were removed from the models.

3.6.2 Linear regressions

The coefficients from univariable linear regressions of each performance metric with every predictor can be found in **Table 9 Table 10** in the appendix. Checking for pairwise interactions of all predictors for all outcomes resulted in no statistically significant interactions. The variables selected into the multiple regression models can be seen in Table 2 and **Table 3**. Surprisingly, the predictors for the same metrics barely overlapped between both systems. The only predictor that was selected equally often for both systems was total counts. HDI appears to be more important for HealthMap than for EIOS performance, as it was selected for three outcomes for HealthMap, but only for one outcome for EIOS. In contrast, global region or latitude was selected for every single variable in EIOS, but not a single time in HealthMap. Remarkably, neither the country's official language nor HealthMap filter language played a big role for the performance of both systems, as they were selected only once each.

Interestingly, increasing the HDI increased sensitivity and prevented fraction but decreased specificity. For instance, sensitivity per outbreak increased on average by 11.5%, and specificity decreased on average by 4.3% with every 0.1 increase in HDI. Likewise, going from tropical to temperate climates in EIOS increased sensitivity per outbreak (0.7% per 1° increase) and prevented fraction, (0.9% per 1° increase) but decreased positive predictive value (-1% per 1° increase). Latitude is a proxy for influenza seasonality, with countries further away from the equator having clear seasonal curves, and countries closer to the equator having more irregular influenza outbreaks. Therefore, the background noise was higher in the latter countries, which complicated outbreak detection.

Moreover, total counts were only selected for sensitivity and timeliness, with higher number of counts increasing the metrics. In HealthMap, an increase of 1 logarithmic unit in counts increased sensitivity per outbreak by 3.9% and the prevented fraction by 5.2%. An increase by the same magnitude in EIOS lead to an increase in sensitivity per week by 0.7% and an increase in prevented fraction by 7.8%. The higher a country ranked on the PFI (i.e. the worse the media freedom), the lower was the specificity in HealthMap (0.3% decrease with every 1 score increase) and the PPV in EIOS (0.4% decrease with every score increase).

For PPV in HealthMap, none of the predictor variables were chosen into the final model, although PPV varied considerably between countries. Specificity had a low number of predictors and low R^2 values in both systems. The adjusted R^2 values ranged from 0.14 to 0.63

and were highest for sensitivity per week in HealthMap and PPV in EIOS. However, these high values can indicate overfitting because the models with the highest R^2 values are using 5 and 7 degrees of freedom on only 24 data points, respectively.

Outcome	Predictor	Category/ Increment	Coefficient [95% CI]	p-value	Adjusted R^2
Sensitivity per outbreak	log(Total counts)	1 log	0.039 [-0.0069 - 0.0845]	0.0920	0.317
	HDI	1 score	1.148 [0.3224 - 1.9733]	0.0087	
Sensitivity per week	log(Total counts)	1 event	0.095 [0.0547 - 0.1358]	0.0001	0.557
	HDI	1 score	0.418 [-0.1667 - 1.0033]	0.1514	
	HealthMap filter language	False	reference		
		True	-0.096 [-0.2465 - 0.0551]	0.2005	
PPV	-				
Specificity	HDI	1 score	-0.429 [-0.9553 - 0.0978]	0.1052	0.141
	PFI	1 score	-0.003 [-0.0056 - -0.0003]	0.0324	
Prevented fraction	log(Total counts)	1 log	0.052 [0.0136 - 0.0896]	0.0101	0.304
	HDI	1 score	0.566 [-0.1196 - 1.2523]	0.1007	

Table 2: Effect of country-specific predictors on HealthMap performance.

Outcome	Predictor	Category/ Increment	Coefficient [95% CI]	p-value	Adjusted R^2
Sensitivity per outbreak	HDI	1 score	1.045 [-0.2206 - 2.310]	0.1004	0.443
	Latitude	1°	0.007 [0.000 - 0.0143]	0.0494	
	PFI	1 score	0.007 [0.0019 - 0.0121]	0.0096	
Sensitivity per week	log(Total counts)	1 log	0.051 [0.0029 - 0.0997]	0.0389	0.358
	Latitude	1°	0.003 [-0.0004 - 0.0071]	0.0792	
PPV	log(Total counts)	1 log	0.037 [-0.0209 - 0.095]	0.1964	0.626
	Global region	Temp. Northern	reference		
		Temp. Southern	0.03 [-0.166 - 0.2262]	0.7507	
		tropical	0.452 [0.2622 - 0.642]	0.0001	
	Latitude	1°	0.01 [0.0046 - 0.0157]	0.0012	
	PFI	1 score	0.004 [-0.0001 - 0.0071]	0.0546	
Specificity	Global region	Temp. Northern	reference		
		Temp. Southern	-0.017 [-0.145 - 0.1101]	0.7783	0.273
		tropical	0.153 [0.0479 - 0.2571]	0.0063	
Prevented fraction	log(Total counts)	1 log	0.078 [-0.0084 - 0.1649]	0.0742	0.489
	Official language	not English	reference		
		English	-0.258 [-0.5566 - 0.0403]	0.0863	
	Latitude	1°	0.009 [0.0035 - 0.0146]	0.0028	

Table 3: Effect of country-specific predictors on EIOS performance.

3.6.3 Logistic regressions

Logistic regression to explore which factors influenced the timeliness of outbreak detection (within 2 weeks after outbreak onset) was problematic because some of the variables showed perfect separation. For example, for the 'Official language' variable, all countries with English being the official language (Australia, UK, and USA) had a timely sensitivity greater than 0 in HealthMap, so fitted probabilities were 1. Similarly, none of the countries with a timely

sensitivity greater than 0 in EIOS are located in the Southern temperate hemisphere. As a consequence, the coefficients and their confidence intervals were inflated. An additional problem was that only 24 data points were available to estimate the coefficients and that the 'English' category of the 'Official language' variable was underrepresented with only 3 countries. For these reasons, logistic regression was not feasible, so the categorical variables with the percentage of successes per stratum are shown (**Table 4**).

HealthMap			**EIOS**		
Official language	**Timely sensitivity**		**Geographical region**	**Timely sensitivity**	
	zero	**> zero**		**zero**	**> zero**
not English	11 (45.8%)	10 (41.7%)	**Temp. Northern hemisphere**	10 (41.7%)	3 (12.5%)
English	0	3 (12.5%)	**Temp. Southern hemisphere**	4 (16.7%)	0
			tropical	5 (20.8%)	2 (8.3%)

Table 4: Variables with perfect separation of timely sensitivity categories in HealthMap and EIOS. *The number of countries in each category are shown in the table, along with the percentages in each category.*

3.6.4 Robustness analysis of variable selection

In order to check the validity of the variable selection, LASSO regressions were performed for every evaluation metric. The variable selection was very similar to selection by AIC, which means that the overall predictor selection process was valid (Table 5). The most striking difference was that LASSO did not select any variables into the specificity model for HealthMap. Additionally, LASSO did not choose any variables to predict timely sensitivity for either system, thus supporting the fact that logistic regression for timely sensitivity is not feasible.

	HealthMap		**EIOS**	
Outcome	**AIC predictors**	**LASSO predictors**	**AIC predictors**	**LASSO predictors**
Sensitivity per outbreak	Total counts HDI	Total counts HDI Global region	HDI Latitude PFI	HDI Latitude PFI Global region
Sensitivity per week	Total counts HDI HM filter language	Total counts HDI	Total counts Latitude	Total counts Latitude
PPV	-	-	total counts Global region Latitude PFI	total counts Global region Latitude PFI
Specificity	HDI PFI	-	Global region	Global region HDI
Prevented fraction	Total counts HDI	Total counts HDI	Total counts Latitude Official language	Total counts Latitude
Timely sensitivity	-	-	-	-

Table 5: Comparison of variable selection with AIC and LASSO.

In a second robustness analysis, the variable selection process was repeated, excluding the three countries with low FluNet data quality (Nigeria, Thailand and Vietnam). Overall, the selected sets of important predictors for each metric were very similar between the full and the reduced datasets. The most striking difference was that in the reduced dataset, three variables were selected as predictors for PPV, which had no significant predictors in the full dataset. The regression coefficients for PPV suggest that the PPV might be lower with higher HDI and HealthMap filter language (regression coefficients in appendix **Table 9**). However, this might also be an artifact because 3 countries with low HDI are deleted. In EIOS, variable selection remained almost unchanged. Overall, it can be concluded that the FluNet data problems did not influence variable selection to a high degree.

	HealthMap		**EIOS**	
Outcome	**Full dataset predictors**	**Reduced dataset predictors**	**Full dataset predictors**	**Reduced dataset predictors**
Sensitivity per outbreak	Total counts HDI	Total counts HDI	HDI Latitude PFI.2018	HDI Latitude PFI.2018
Sensitivity per week	Total counts HDI HealthMap filter language	Total counts	Total counts Latitude	Total counts Latitude
PPV	-	Official language HDI.2018 HealthMap filter language	Total counts Global region Latitude PFI	Global region Latitude PFI
Specificity	HDI PFI	HDI Official language Total counts	Global region	
Prevented fraction	Total counts HDI	Total counts HDI	Total counts Official language Latitude	Total counts Official language Latitude

Table 6: Comparison of variable selection between the full dataset and the reduced dataset. *In the reduced dataset, countries with low FluNet data quality (Nigeria, Thailand, and Vietnam) were excluded.*

4 Discussion

4.1 Principal findings

4.1.1 How did the systems perform?

This study assessed one way in which EBS systems can be used to detect disease outbreaks. In this work, in order to formally evaluate the performance of HealthMap and EIOS, their ability to detect seasonal influenza outbreaks in 24 countries worldwide was compared with a gold standard based on FluNet. Outbreaks were detected by Bayesian change point (BCP) analysis both in the EBS event data and in the lab-confirmed cases from FluNet. It is important to note that all analyses were done retrospectively on the complete datasets, and not prospectively like in real-time surveillance. Performance metrics varied widely between the 24 countries and the two systems.

Sensitivity per outbreak was used to analyze the crude detection of outbreaks and was over 75% for most countries. EIOS detected all outbreaks in 7 countries, but also detected none of the outbreaks in one country, namely Costa Rica. Specificity was the evaluation metric in which both systems were found to have the most similar values across countries. In contrast, positive predictive value (PPV) and timeliness differed greatly between countries and systems. While all the above metrics measure some characteristic of the EBS system, they are only meaningful if the systems' users are interested in improving one metric at a time or focusing on one metric only. For example, users might want to have a system alerting with high specificity if the resources to respond to alerts are constrained. In contrast, they might want high sensitivity if other information suggests that the likelihood of a disease outbreak is high. However, to my knowledge, the user perspectives on the functionalities of EBS systems have not been described in the literature.

In reality, users would want to work with a system that detects outbreaks both timely and accurately. This is why composite metrics were evaluated, namely accuracy and F1. Accuracy combines sensitivity and specificity, and the majority of countries had an accuracy between 45% and 75%. The F1 score is more often used in information retrieval, machine learning, and natural language processing than in diagnostic testing. However, since it combines sensitivity and positive predictive value into one measure, it is also a useful evaluation metric for the performance of HealthMap and EIOS. The F1 metric showed more variability across countries. While EIOS rarely achieved an F1 score over 50%, HealthMap had a score of over 50% in almost half of the countries. In summary, HealthMap showed a consistently higher accuracy and F1 score than EIOS. What was surprising was the high accuracy of HealthMap in Ecuador and Germany, since the events from these countries are only spikes of few counts. This

provides even more evidence than the comparison of mean evaluation metrics that spikes at the right time are useful.

Accuracy and the F1 score are valuable to obtain a combined picture of sensitivity and specificity or PPV, but not about timeliness. Therefore timeliness and sensitivity were combined into a timely sensitivity metric, which is the most interesting metric for evaluation an EBS system. Timely sensitivity provides an estimation of the systems' ability to detect infectious diseases outbreaks in a timely manner, i.e. before traditional surveillance systems, which is the main reason why EBS was developed. The aggregation of daily event counts into weekly counts contradicts this timeliness idea, but otherwise event counts would have been too low for a meaningful analysis in some countries, and the comparability with FluNet would not have been given. The ability of systems to timely detect outbreaks was disappointing: EIOS had only 5/24 countries in which the timely detection of outbreaks was over 0%. The highest number for any country for HealthMap was 2/7 outbreaks detected in a timely manner in three countries, and in 10/24 countries, one outbreak was detected on time. A first conclusion is therefore that caution should be exercised when using the evaluated systems alone in the manner used in this study for infectious disease surveillance, as an analysis of event frequency has shown non-satisfactory results for influenza surveillance. HealthMap and EIOS might prove useful in combination with other sources such as social media or environmental data, however, or they may also give better results if the online media are manually filtered and examined qualitatively (e.g. by human analysts).

4.1.2 Which factors influenced system performance?

As many determinants affect the performance of outbreak detection in EBS, discovering how these factors influence the detection ability can help to improve these systems. An analysis of determinants of performance also helps to understand in which contexts and why using event frequency is useful. Therefore, the relationship of the detection performance with various country-specific factors was examined in regressions.

Data abundance, measured by total counts per country, was a factor influencing both HealthMap and EIOS performance. It was selected as an influential variable for sensitivity, PPV, and timeliness, with higher count numbers leading to an increase in all these metrics. The human development index (HDI) was more influential for HealthMap than EIOS, as it was chosen for 4 out of 6 metrics in HealthMap models and only once for EIOS models. An increase in HDI improved sensitivity and timeliness, but decreased the positive predictive value. The HDI was strongly correlated with total number of internet users, but surprisingly not with HealthMap event numbers and only moderately with EIOS event numbers. So the influence of HDI on the performance metrics was not mediated through event numbers. In contrast, HealthMap filter language was only chosen as a predictor once, and highly correlated with

HealthMap event counts. Likewise, English as a country's official language increased event numbers. This is in line with a previous observation that HealthMap received its vast majority of events from English sources [37]. Due to their high correlation with total counts, the language variables were probably acting on sensitivity only indirectly. The predictor which was chosen most often for EIOS performance was a country's geographic location, either as global region or latitude. These variables were associated with influenza seasonality, but also with EIOS event counts and HDI. For countries further away from the equator, EIOS showed higher sensitivity, timeliness and PPV, but reduced specificity. However, there was no obvious threshold effect for any of the predictors.

Looking at the R^2 values of predictors in univariable regressions, total counts had on average the largest influence on HealthMap performance. In EIOS, the results were less obvious, with geographical location having the largest influence on timeliness, specificity, and PPV, and total counts playing a less important role.

Even after regressing the performance metrics on the predictors, a high degree of variability between the countries was still present. Especially the specificity models had very low R^2 values, indicating that there are other sources of variability that have not yet been identified. Even more extreme, no suitable predictors for HealthMap PPV could be found. In contrast, the models for HealthMap sensitivity by week and EIOS PPV showed high R^2 values. This indicates potential overfitting, as from only 24 degrees of freedom, the models took 5 and 7, respectively. The same problem was present for the logistic regression with timely sensitivity as outcome. The very low number of data points, even lower number of 'successes' in EIOS, and low counts in certain categories led to instable and unreliable estimates, so no meaningful logistic regression was feasible.

4.1.3 Differences between HealthMap and EIOS

Overall, HealthMap performed consistently better than EIOS, based on the mean of all performance metrics. This highlights the differences in conceptual design and functionality of the two systems. While EIOS data could only be collected over 2 years, HealthMap contributed 6.5 years of data. This shows that more events do not necessarily lead to a better performance, and can lead to more residual noise. In a comparison study of three EBS systems, Lyon et al. noted HealthMap had less events detected than other evaluated systems, which were BioCaster and EpiSPIDER [37]. However, HealthMap had far more "quality" reports than the other two systems. The authors hypothesized that HealthMap's data are less noisy and more informative because a significant percentage of HealthMap's reports came from its community of users.

The conceptual differences between HealthMap and EIOS are shown even clearer by the fact that a different set of predictors was chosen for both systems. Whereas HDI played an important role for HealthMap performance, EIOS was more influenced by a country's geographical location.

HealthMap data structure was different from EIOS in another aspect: In lower count countries, spikes of a few events per week were visible. Introducing a count cutoff of minimum 5 counts for outbreak detection increased sensitivity per week and timeliness, but did not change sensitivity per outbreak, and even decreased PPV and specificity. As it decreased the composite metrics, too, it was concluded that a count limit for raising an alarm does not improve HealthMap performance. However, it is questionable if in practice an alarm based on only a few events will be raised by the system's users. Moreover, these few events might go unnoticed, and so outbreaks will be missed.

Since EIOS uses HealthMap alerts in addition to other online data sources, HealthMap represents the more conservative data source between the two. Unfortunately, it was not possible to identify which input was contributed to EIOS by HealthMap and which by other sources because this information was not supplied in the data. However, it is clear that at least some of these sources contribute noisy information, as EIOS generally performed worse than HealthMap.

4.2 Comparison with other studies

Scientists have long stressed the need for analyzing and quantifying the output of biosurveillance systems [23]. In one of the few available studies about the differential performance of EBS systems, Barboza et al. analyzed six biosurveillance systems and compared their detection rate, PPV, F1, sensitivity and timeliness of detection of H5N1 outbreaks [68]. In a qualitative survey, end users reported using HealthMap as a complementary source of biosurveillance. In the quantitative section, HealthMap detected outbreaks on average 12 days before the gold standard, so a much earlier outbreak detection than in this work. However, the detection rate was only 43% (compared to an average outbreak sensitivity across countries of 75% in this work), and a PPV of 12% (compared to 66% in this work). These striking differences are likely to stem from discrepancies in study design: Barboza et al. did not attempt an outbreak detection based on event counts, but counted the first event relating to a respective H5N1 outbreak as a true positive. This approach is debatable, as it is unlikely that an outbreak alarm is raised based on only one event. Additionally, the authors of this study used WHO reports on H5N1 as a gold standard. These reports require official notification by a national authority, a process that can take time, and are thus not very accurate in timing [68].

In another study, Barboza et al. compared the same six systems for their ability to detect various infectious disease outbreaks, while using the weekly international epidemiological bulletin from the French Institute for Public Health Surveillance (InVS) as a gold standard [63]. Additionally, they identified factors which influenced system performance, such as types of system, languages, regions of occurrence, and types of infectious disease. However, they did not compare country-specific characteristics. Interestingly, they stress the importance of developing a common biosurveillance tool for aggregating system outputs, which was later realized with the development of EIOS, with Dr. Barboza as project lead. In a previous study, they had indeed constructed a virtual combined system from six sources and assessed its performance [68]. This combined system achieved a 93% detection rate of human H5N1 outbreaks, but only a 7% PPV and a 13% F1 score. Compared to that, EIOS had a mean sensitivity per outbreak of 73%, a PPV of 60%, and an F1 score of 34% in this work, so the actual aggregate system is less sensitive, but has a higher PPV. Nevertheless, this example illustrates that event data get messier when combined, as there is more available information, but also more noise.

Hoen et al. determined the sensitivity and specificity of HealthMap in detecting new Dengue virus (DENV) circulation in previously DENV-non-endemic regions in Latin America, using the CDC's Yellow Book as a reference [105]. HealthMap's timeliness far outperformed the traditional system, which is most likely due to limitations of traditional system using passive case reports. The overall sensitivity of HealthMap for detecting DENV outbreaks was 74%, and the specificity was 85%. While these values are more similar to the ones found in this work, all the above examples clearly demonstrate the lack of a reliable gold standard for comparing the performance of EBS systems, an issue which has been criticized by other authors as well [23], [67]. This not only leads to differential results in evaluation metrics for multiple systems, but also creates problems when comparing the same system across diseases and regions.

When analyzing factors influencing influenza detection in South America with Google Flu Trends (GFT), Pollett et al. found that countries further away from the equator had a better correlation with FluNet data, probably due to more regular seasonality [106]. While this observation was replicated with EIOS events in this work, it might have been an artifact of the GFT algorithm in Pollett's study. GFT has been shown to fit structurally unrelated search terms, thus just predicting seasonality and not influenza counts [107].

Nevertheless, this study highlights another difference which complicates EBS performance comparisons across studies: The metric to compare novel biosurveillance approaches with gold standards which is used in a lot of studies is the Pearson correlation coefficient [18], [59], [108]–[110]. While it allows for easy computing of a comparison metric, it is highly susceptible to influential data points and assumes that infectious disease counts from adjacent points in a

time series are independent [67]. Moreover, correlation coefficients are biased by low-frequency patterns in the data [111]. Furthermore, correlation coefficients are less meaningful from a user perspective than timeliness, sensitivity, and specificity. Therefore, more user-centered evaluation metrics have been chosen for this work.

Because of the restrictions of choosing an outbreak detection method (see section 2.2.1), several established methods had to be discarded: Serfling regression models seasonality and requires knowledge of a “baseline” period [81], [82]. Threshold methods developed by Cowling [83] or Neuzil [84] are impractical because choosing thresholds is an arbitrary process and thresholds would have to be adapted to each country separately. Moreover, setting an appropriate threshold requires pre-existing knowledge of epidemic and non-epidemic periods. Poisson or linear regression methods adjust for seasonality and require long training periods [82], [85], as does the widely used Farrington algorithm [86]. EARS algorithms are very basic and are heavily dependent on the choice of length of the baseline period [87], [88]. The *outbreakP* method was specifically designed to detect influenza outbreaks [89], but since it is limited to detecting only one outbreak in a time series, it was not applicable to the data sets in this project. The exponentially weighted moving average (EWMA) method requires individual baseline calculation for each country [90]. Lastly, times series models require years of training data and modeling of seasonality [91]. Machine learning (ML) methods have so far been applied to infectious disease outbreak detection mostly to combine multiple large datasets such as social media data, weather data, and traditional surveillance data [92]–[94]. Therefore, no additional use of ML methods was found for this study. Bayesian change point (BCP) analysis was used as it is not based on arbitrary thresholds or baselines, nor does it require a training period or seasonality indicators. Furthermore, the Bayesian framework allows to deal with uncertainties in estimates in an elegant way.

4.3 Strengths

This work is the first study rigorously evaluating EBS against a clear gold standard on a global scale, permitting country-specific external factors influencing performance to be identified. Moreover, the focus of biosurveillance was broadened from mostly rich Western countries to a worldwide data analysis through this study. Multiple outbreak patterns of seasonal influenza could be studied by including countries from 15 influenza transmission zones into this study.

Another strength was that the method of outbreak detection used was easy, intuitive, and reproducible. BCP analysis is not traditionally used in infectious disease outbreak detection, but it performed well in detecting the beginning of rising curves. Shmueli and Burkom remark: "The task [of outbreak detection] is one of anomaly detection rather than signature identification."[112] , and this is precisely what BCP has been developed for. Since the same

method was applied to all three datasets, biased conclusions based on differential outbreak detection were avoided.

As discussed before, the choice of gold standard greatly influences the results of performance evaluations. With FluNet, a very accurate gold standard was chosen for this study, which has been used for studying influenza epidemic patterns worldwide [47], [75], [106]. The lab-confirmed influenza cases are reported with about 2 weeks delay, but assigned the correct date, unlike reports from other gold standards, which rely on passive case reporting.

Another strength of the study is the robustness of the influential variable selection. Variable selection through LASSO resulted in almost the same predictors as through the AIC criterion. The subsets of influential variables chosen on the full dataset and on the dataset excluding countries with low FluNet data quality were very similar as well. This result is even more remarkable regarding the fact that the whole sample consisted of only 24 countries.

4.4 Limitations

However, the study also had limitations, which will be discussed in the next paragraphs. First, this work was focused on only one way EBS could be used for early detection. There may be other approaches, which draw on 'human-in-the-loop' or other data, which would potentially perform better.

Second, gold standard data were not labeled according to epidemic periods, so influenza outbreaks had to be detected with the same method as in the EBS data. While this approach was valid for most countries and the same bias – if any – would apply to all three datasets, FluNet outbreak detection was not optimal in Nigeria, Thailand, and Vietnam due to the low number of reported cases and the inherently more irregular influenza activity. Therefore, the labeling of epidemic and non-epidemic periods in these countries might not have been reliable. However, in a robustness analysis, the selection of system performance predictors was found to be very similar between the whole dataset and the reduced dataset without Nigeria, Thailand, and Vietnam.

FluNet data suffer from three additional shortcomings: Firstly, the counts of influenza-infected people stem from people who have sought healthcare and have had a swab taken, thus underestimating the total amount of influenza activity in a country. In other words, the sensitivity of FluNet is not very high. Secondly, surveillance activities are not uniform across countries, so the number of swabs taken is highly dependent on healthcare system capacity with a lot of testing in developed countries and lower testing efforts in developing countries. Additionally, some countries show reduced interest in post-peak activities [113]. For instance, France only reports influenza cases from the beginning of October to the beginning of May. Thirdly, FluNet publishes the counts with a time lag of at least one week in developed countries and at least

two weeks in other countries. Moreover, numbers can be revised after the initial upload. Therefore, these data cannot be used for timely detection of influenza outbreaks. Nevertheless, they are highly specific and have a high positive predictive value because the number of false negative laboratory tests is very low. Therefore, lab-confirmed case counts accurately reflect the start and end points of epidemic periods, which are the outcomes of interest.

Third, the event-based data sources may have limitations: HealthMap events were very sparse in some countries, so that every event was classified as an outbreak. Raising an alarm for only a single instance of influenza reporting is not practical in real-life disease surveillance. EIOS had collected many more events than HealthMap, but only over a span of 2 years, and the data were much noisier. Thus, BCP analysis did not work optimally, and did not recognize some of the spikes or lower peaks as outbreaks. Unfortunately, it was not possible to separate the EIOS input according to source in order to identify which sources contribute signal and which create only noise. Another related issue is that EIOS usually mentions several countries per event, either because an event really affects multiple countries or, more problematically, because of structural defaults. For example, all events reported by GPHIN are classified as mentioning Canada, even if the disease is located in another part of the world. Furthermore, EIOS erroneously assigns events to broad disease categories. For instance, a lot of COVID-19 reports on the EIOS platform are also classified as “influenza not specified” (observation from May 21st, 2020, data not shown). While SARS-CoV-2 had not yet emerged at the time of data collection, other diseases might have been falsely categorized as influenza. A qualitative analysis of all disease categories contained in the reports was not conducted.

Another limitation of EBS is that multiple influenza strains and respiratory infections are aggregated into one category. Such aggregation poses difficulties in modeling because different strains of the flu exhibit different seasonal characteristics. As the EBS data in this study are aggregated per country, they do not capture any regional diversity within countries. This is especially problematic for noncontiguous land masses and large countries spanning diverse climatic regions such as China, Brazil, and the USA, with different epidemic properties of influenza [113]. In the analysis of influential performance predictors, latitude was just assigned as one value in the midpoint of a country. As HealthMap provides the putative latitude and longitude of each event, it would be possible to analyze events at a more granular spatial resolution. However, to my knowledge the accuracy of this geographical allocation has never been examined.

Moreover, all EBS data face some inherent challenges (see section 1.1.3). Some of these, such as false positive signals, could be observed in both EBS datasets. Dependency on internet coverage was identified only in HealthMap data because they were dependent on the

HDI and thus on percentage of internet users per country. A language dependency could not be detected in either dataset according to the selected predictors, although language variables might have been missed because of the small sample size and the low numbers of countries in the language categories. Frontloading, i.e. the uneven distribution of events to the start of epidemics, was not obvious from the visual correlations. Even if it occurred, it could not have influenced the timeliness and timely sensitivity variables. Crowding-out phenomena, that is, certain diseases and epidemics temporarily suppressing the reporting of others, could have happened during the study period. Certainly, the current SARS-CoV-2 pandemic right now and the worldwide H1N1 outbreak in 2009 temporarily led to less media coverage of other diseases [114]. Monitoring news articles can also be affected by significant day-of-the-week effects, with less reporting on weekends and holidays. Since EBS data were aggregated per week, this was not an issue in this study. However, by aggregating counts per week, the potential benefit of daily available data for timely outbreak detection was lost.

Fourth, a significant limitation of this study is the small sample size. Since data was only available for 24 countries, the regressions had to be performed with a small number of degrees of freedom and low numbers of countries per category in categorical variables. Therefore, confidence intervals were very wide, and perfect separation occurred in the logistic regressions, so that coefficients were no longer interpretable.

Fifth, BCP was used as the method to detect influenza outbreaks despite probably not meeting the distributional assumption of Normality. This is a limitation because BCP might have missed or wrongly classified some change points in the datasets. However, as explained in section 2.2.2, BCP should be able to separate counts during epidemic and non-epidemic periods due to their very different means. Another inherent disadvantage of BCP is that it cannot be used as a technique to analyze real-time data because it relies on calculating the means of each block. In this study, BCP was used to retrospectively detect influenza outbreaks, and so the outbreak detection algorithm cannot be directly applied to prospective surveillance. Barboza et al. found in a study that the prospective sensitivity of EBS systems was 17% lower than the retrospective sensitivity [68], which highlights even more the differences between retrospective and prospective disease surveillance. Since BCP was applied retrospectively on all datasets, the evaluation results likely overestimate the true performance of EBS systems.

4.5 Future research

This study has laid the foundation for evaluating the performance of HealthMap and EIOS. Since EIOS data was limited to 2 years, it would be interesting to see if the evaluation metrics remain stable with a larger dataset. EIOS is still in its development phase, hence its performance could be greatly improved by updates in the near future. The analysis of important predictors of performance was only exploratory and hypothesis-generating due to the small

sample size. A study including more countries would result in a more definitive and relevant set of influential variables.

Another aspect that warrants more research is the exploration of EIOS sources in order to identify which contribute valuable and accurate input and which generate unspecific noise. Timely sensitivity was the most important evaluation metric in this study, yet especially EIOS failed to detect outbreaks on time. Therefore, more research should be done to increase EIOS' timely disease detection capabilities. Another feature which would be helpful for users of EBS is the automatic extraction of case counts from articles, which requires more research in natural language processing.

Moreover, input from HealthMap and EIOS could be used along with other data sources such as IBS, meteorological data, social media, and Google or Wikipedia searches to forecast influenza counts in various countries. It would be interesting to see if EBS systems improve forecasts, as EIOS has never been used for forecasting, or at least no such studies have been published so far.

4.6 Conclusions and recommendations

To my knowledge, this is one of the first studies to rigorously evaluate the performance of HealthMap and EIOS on a global scale. The study was done across 24 countries and assessed their ability to detect influenza outbreaks. To divide the two EBS data sets and the gold standard into epidemic and non-epidemic periods, Bayesian change point analysis was used. While outbreak detection and specificity were generally high in both systems, positive predictive value and timeliness varied considerably across countries and systems. In contrast, both systems failed to detect outbreaks in a timely manner in many countries and only detected few outbreaks on time in the others.

This means that event counts alone are not a very good method to detect influenza outbreaks in a timely manner, and so any useful analysis has to combine event frequency from EBS with a contextual analysis, extracted either from media reports or obtained from other systems, and historical or environmental data. The countries for which HealthMap worked best were the USA, and EIOS had the best performance in the UK and France.

Identification of influential country-specific factors on performance revealed that influenza outbreak detection by EBS systems is better in countries with higher HDI, countries further away from the equator, and in countries for which the systems had generated more events. However, no threshold effects indicating that a system works well in a country with a certain number of events or a certain geographical location have been found. Therefore, and in order to avoid over-fitting to the disease and the limited data in some countries, no recommendation

will be made in which countries to use HealthMap or EIOS. Two sensitivity analyses confirmed a robust predictor selection.

Feedback from the end users will be essential in order to improve certain characteristics of the systems, such as timeliness or specificity. Additional research that should be done on EIOS to determine how diseases can be classified into more meaningful categories and how countries can be properly allocated in the events. A relevancy score which is automatically generated for each article could help users to identify important events, and acquiring more sources to increase events pertaining to low-count countries will improve overall system performance.

5 Conclusion on my experience as a professional

The most important experience from the internship was working independently on the same project for a long time. Although I have gained experience in working on longer projects before because I did another Master's degree, I did not work on an epidemiological project. Since my project was not very team-oriented, I was mostly working by myself, but it was still good to see the other students and researchers in my lab every day. After the lockdown due to Covid-19, I had to work from home. During this time, it was not easy for me to stay focused and sometimes I really felt overwhelmed. What helped me was breaking down the tasks into small chunks so that I had a task for every day. This helped me achieve major breakthroughs like finding an outbreak detection method, getting this method to work on my data, the calculation of performance metrics, and the identification of influential factors on performance.

Another interesting experience was having two supervisors on two different continents with different backgrounds. The discussions with them over Sykpe or Zoom were very stimulating and important for the progress of my work. It was interesting to see the scientific collaboration between them.

Additionally, through this internship I developed new skills, especially coding in R, learning new statistical methods like BCP, being persistent on one project for a long time, literature research, and having discussions through Skype.

In general, I am very happy with the outcome of this project, as I achieved the goals that I wanted to achieve and prepared the contents for a scientific paper. The only thing that did not go as planned is that I did not have enough time to do influenza modeling with EBS as one of the input sources.

6 References

[1] World Health Organization, “WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases,” 2000.

[2] World Health Organization, “International Health Regulations 3rd edition,” 2005.

[3] World Health Organization, “Implementation of Early Warning and Response with a focus on Event-Based Surveillance,” *WHO*, pp. 1–64, 2014.

[4] C. Abat, H. Chaudet, J. M. Rolain, P. Colson, and D. Raoult, “Traditional and syndromic surveillance of infectious diseases and pathogens,” *International Journal of Infectious Diseases*, vol. 48. Elsevier, pp. 22–28, 01-Jul-2016.

[5] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009.

[6] C. K. Cheng, E. H. Lau, D. K. Ip, A. S. Yeung, L. M. Ho, and B. J. Cowling, “A profile of the online dissemination of national influenza surveillance data,” *BMC Public Health*, vol. 9, no. 1, p. 339, Dec. 2009.

[7] S. S. Morse, “Public Health Surveillance and Infectious Disease Detection,” *Biosecurity Bioterrorism Biodefense Strateg. Pract. Sci.*, vol. 10, no. 1, pp. 6–16, Mar. 2012.

[8] K. J. Henning, “What is syndromic surveillance?,” *MMWR. Morb. Mortal. Wkly. Rep.*, vol. 53 Suppl, pp. 5–11, 2004.

[9] M. D. Lewis *et al.*, “Disease outbreak detection system using syndromic data in the greater Washington DC area,” *Am. J. Prev. Med.*, vol. 23, no. 3, pp. 180–186, Oct. 2002.

[10] E. H. Chan, R. Tamblyn, K. M. L. Charland, and D. L. Buckeridge, “Outpatient physician billing data for age and setting specific syndromic surveillance of influenza-like illnesses,” *J. Biomed. Inform.*, vol. 44, no. 2, pp. 221–228, Apr. 2011.

[11] E. Vergu *et al.*, “Medication Sales and Syndromic Surveillance, France,” *Emerg. Infect. Dis.*, vol. 12, no. 3, pp. 416–421, Mar. 2006.

[12] E. H. Chan, V. Sahai, C. Conrad, and J. S. Brownstein, “Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance,” *PLoS Negl. Trop. Dis.*, vol. 5, no. 5, p. e1206, May 2011.

[13] R. Desai, B. A. Lopman, Y. Shimshoni, J. P. Harris, M. M. Patel, and U. D. Parashar, “Use of Internet Search Data to Monitor Impact of Rotavirus Vaccination in the United States,” *Clin. Infect. Dis.*, vol. 54, no. 9, pp. e115–e118, May 2012.

[14] R. Desai *et al.*, “Norovirus Disease Surveillance Using Google Internet Query Share Data,” *Clin. Infect. Dis.*, vol. 55, no. 8, pp. e75–e78, Oct. 2012.

[15] V. M. Dukic, M. Z. David, and D. S. Lauderdale, “Internet Queries and Methicillin-Resistant *Staphylococcus aureus* Surveillance,” *Emerg. Infect. Dis.*, vol. 17, no. 6, pp. 1068–1070, Jun. 2011.

[16] Xichuan Zhou, Jieping Ye, and Yujie Feng, “Tuberculosis Surveillance by Analyzing Google Trends,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 8, pp. 2247–2254, Aug. 2011.

[17] A. Alessa and M. Faezipour, “A review of influenza detection and prediction through social networking sites.,” *Theor. Biol. Med. Model.*, vol. 15, no. 1, p. 2, 2018.

[18] D. A. Broniatowski, M. J. Paul, and M. Dredze, “National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic,” *PLoS One*, vol. 8, no. 12, p. e83672, Dec. 2013.

[19] M. J. Paul, M. Dredze, and D. Broniatowski, “Twitter improves influenza forecasting.,” *PLoS Curr.*, vol. 6, Oct. 2014.

[20] J. Gomide *et al.*, “Dengue surveillance based on a computational model of spatio-temporal

locality of Twitter," in *Proceedings of the 3rd International Web Science Conference, WebSci 2011*, 2011, pp. 1–8.

[21] World Health Organization, "Early detection, assessment and response to acute public health events," 2014.

[22] J. O'Shea, "Digital disease detection: A systematic review of event-based internet biosurveillance systems," *Int. J. Med. Inform.*, vol. 101, pp. 15–22, May 2017.

[23] D. Hartley *et al.*, "The landscape of international event-based biosurveillance," *Emerg. Health Threats J.*, vol. 3, no. 1, p. 7096, 2010.

[24] L. C. Madoff and J. P. Woodall, "The Internet and the Global Monitoring of Emerging Diseases: Lessons from the First 10 Years of ProMED-mail," *Arch. Med. Res.*, vol. 36, no. 6, pp. 724–730, Nov. 2005.

[25] E. Mykhalovskiy and L. Weir, "The Global Public Health Intelligence Network and Early Warning Outbreak Detection," *Rev. Can. Sante Publique*, vol. 97, no. 1, pp. 42–44, 2005.

[26] A. Mawudeki and M. Blench, "Global Public Health Intelligence Network (GPHIN)," 2006.

[27] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports," *J. Am. Med. Informatics Assoc.*, vol. 15, no. 2, pp. 150–157, Mar. 2008.

[28] N. Collier *et al.*, "BioCaster: detecting public health rumors with a Web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, Dec. 2008.

[29] J. Linge *et al.*, "MedISys: Medical Information System.," in *Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks*, E. Asimakopoulou and N. Bessis, Eds. IGI Global, 2010, pp. 131–142.

[30] P. Abdelmalik, E. Peron, J. Schnitzler, J. Fontaine, E. Elfenkampera, and P. Barboza, "The epidemic intelligence from open sources initiative: a collaboration to harmonize and standardize early detection and epidemic intelligence among public health organizations.," *Wkly. Epidemiol. Rec.*, vol. 93, no. 20, pp. 267–269, 2018.

[31] J. S. Brownstein *et al.*, "Combining Participatory Influenza Surveillance with Modeling and Forecasting: Three Alternative Approaches," *JMIR Public Heal. Surveill.*, vol. 3, no. 4, 2017.

[32] M. Dion, P. AbdelMalik, and A. Mawudeku, "Big Data and the Global Public Health Intelligence Network (GPHIN).," *Can. Commun. Dis. Rep.*, vol. 41, no. 9, pp. 209–214, Sep. 2015.

[33] L. HOSSAIN, D. KAM, F. KONG, R. T. WIGAND, and T. BOSSOMAIER, "Social media in Ebola outbreak," *Epidemiol. Infect.*, vol. 144, no. 10, pp. 2136–2143, Jul. 2016.

[34] M. Odlum and S. Yoon, "What can we learn about the Ebola outbreak from tweets?," *Am. J. Infect. Control*, vol. 43, no. 6, pp. 563–71, Jun. 2015.

[35] D. M. Hartley *et al.*, "An overview of Internet biosurveillance," *Clin. Microbiol. Infect.*, vol. 19, no. 11, pp. 1006–1013, Nov. 2013.

[36] R. Bernard, G. Bowsher, C. Milner, P. Boyle, P. Patel, and R. Sullivan, "Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks.," *J. Public Health (Bangkok).*, vol. 26, no. 5, pp. 509–514, 2018.

[37] A. Lyon, M. Nunn, G. Grossel, and M. Burgman, "Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap," *Transbound. Emerg. Dis.*, vol. 59, no. 3, pp. 223–232, Jun. 2012.

[38] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project," *PLoS Med.*, vol. 5, no. 7, p. e151, Jul. 2008.

[39] L. Held and M. Paul, "Modeling seasonality in space-time infectious disease surveillance data," *Biometrical J.*, vol. 54, no. 6, pp. 824–843, Nov. 2012.

[40] J. P. Linge *et al.*, “Internet Surveillance Systems for Early Alerting of Health Threats.”

[41] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V Marathe, “A systematic review of studies on forecasting the dynamics of influenza outbreaks,” *Influenza Other Respi. Viruses*, vol. 8, no. 3, pp. 309–316, May 2014.

[42] World Health Organization, “WHO | FluNet,” *WHO*, 2017. [Online]. Available: https://www.who.int/influenza/gisrs_laboratory/flunet/en/. [Accessed: 27-Jul-2019].

[43] World Health Organization, “WHO Fact Sheet on Seasonal Influenza,” 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). [Accessed: 20-May-2019].

[44] S. Modrow, D. Falke, U. Truyen, and H. Schätzl, *Molecular Virology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[45] A. Budd *et al.*, “Surveillance Manual - Chapter 6: Influenza.”

[46] World Health Organization, “Influenza transmission zones,” 2018.

[47] L. P. Newman, N. Bhat, J. A. Fleming, and K. M. Neuzil, “Global influenza seasonality to inform country-level vaccine programs: An analysis of WHO FluNet influenza surveillance data between 2011 and 2016,” *PLoS One*, vol. 13, no. 2, p. e0193263, Feb. 2018.

[48] R. Garten *et al.*, “Update: Influenza activity in the united states during the 2017–18 season and composition of the 2018–19 influenza vaccine,” *Morb. Mortal. Wkly. Rep.*, vol. 67, no. 22, pp. 634–642, Jun. 2018.

[49] M. Biggerstaff *et al.*, “Results from the second year of a collaborative effort to forecast influenza seasons in the United States,” *Epidemics*, vol. 24, pp. 26–33, Sep. 2018.

[50] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, “Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance,” *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004513, Oct. 2015.

[51] J.-P. George, D. Shaman, J. A. Chitale, and R. A. Mckenzie, “Influenza Forecasting in Human Populations: A Scoping Review,” *PLoS One*, vol. 9, no. 4, p. 94130, 2014.

[52] N. G. Reich *et al.*, “A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 8, pp. 3146–3154, Feb. 2019.

[53] “Flu Trends: Estimate of percentage of people with flu like symptoms.” [Online]. Available: https://www.healthmap.org/flutrends/. [Accessed: 21-May-2020].

[54] M. Santillana *et al.*, “Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance,” *Sci. Rep.*, vol. 6, no. 1, pp. 1–8, May 2016.

[55] P. Chakraborty *et al.*, “Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions,” in *Proceedings of the 2014 SIAM International Conference on Data Mining*, 2014, pp. 262–270.

[56] T. Rekatsinas *et al.*, “SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources.”

[57] S. Bhatia *et al.*, “Using Digital Surveillance Tools for Near Real-Time Mapping of the Risk of International Infectious Disease Spread: Ebola as a Case Study,” *medRxiv*, p. 19011940, Nov. 2019.

[58] N. Hossain and M. Househ, “Using HealthMap to Analyse Middle East Respiratory Syndrome (MERS) Data.,” *Stud. Health Technol. Inform.*, vol. 226, pp. 213–216, 2016.

[59] S. F. McGough, J. S. Brownstein, J. B. Hawkins, and M. Santillana, “Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data,” *PLoS Negl. Trop. Dis.*, vol. 11, no. 1, p. e0005295, Jan. 2017.

[60] M. S. Majumder, M. Santillana, S. R. Mekaru, D. P. McGinnis, K. Khan, and J. S. Brownstein, “Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak,” *JMIR Public Heal. Surveill.*, vol. 2, no. 1, p. e30, Jun. 2016.

[61] W. Naudé, “Artificial Intelligence against COVID-19: An Early Review,” 1311.

[62] World Health Organization, “Early detection, verification, assessment and communication.” [Online]. Available: https://www.who.int/eios. [Accessed: 22-May-2020].

[63] P. Barboza *et al.*, “Factors Influencing Performance of Internet-Based Biosurveillance Systems Used in Epidemic Intelligence for Early Detection of Infectious Diseases Outbreaks,” *PLoS One*, vol. 9, no. 3, p. e90536, Mar. 2014.

[64] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, “HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports,” *J. Am. Med. Informatics Assoc.*, vol. 15, no. 2, pp. 150–157, Mar. 2008.

[65] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, “Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project,” *PLoS Medicine*, vol. 5, no. 7. pp. 1019–1024, Jul-2008.

[66] D. L. Heymann and G. Rodier, “Brown Journal of World Affairs SARS: A Global Response to an International Threat,” 2004.

[67] S. Pollett, B. M. Althouse, B. Forshey, G. W. Rutherford, and R. G. Jarman, “Internet-based biosurveillance methods for vector-borne diseases: Are they novel public health tools or just novelties?,” *PLoS Neglected Tropical Diseases*, vol. 11, no. 11. Public Library of Science, 30-Nov-2017.

[68] P. Barboza *et al.*, “Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events,” *PLoS One*, vol. 8, no. 3, p. e57252, 2013.

[69] J. D. Sharpe, R. S. Hopkins, R. L. Cook, and C. W. Striley, “Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: A Comparative Analysis,” *JMIR Public Heal. Surveill.*, vol. 2, no. 2, p. e161, Oct. 2016.

[70] S. Yousefinaghani, R. Dara, Z. Poljak, T. M. Bernardo, and S. Sharif, “The Assessment of Twitter’s Potential for Outbreak Detection: Avian Influenza Case Study,” *Sci. Rep.*, vol. 9, no. 1, p. 18147, Dec. 2019.

[71] S. Yang, M. Santillana, J. S. Brownstein, J. Gray, S. Richardson, and S. C. Kou, “Using electronic health records and Internet search information for accurate influenza forecasting,” *BMC Infect. Dis.*, vol. 17, no. 1, p. 332, Dec. 2017.

[72] W. Jia *et al.*, “Integrating Multiple Data Sources and Learning Models to Predict Infectious Diseases in China.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2019, pp. 680–685, 2019.

[73] S. J. Yan, A. A. Chughtai, and C. R. Macintyre, “Utility and potential of rapid epidemic intelligence from internet-based sources,” *Int. J. Infect. Dis.*, vol. 63, pp. 77–87, Oct. 2017.

[74] World Health Organization, “Global Epidemiological Surveillance Standards for Influenza,” Geneva, 2013.

[75] E. Azziz Baumgartner *et al.*, “Seasonality, Timing, and Climate Drivers of Influenza Activity Worldwide,” *J. Infect. Dis.*, vol. 206, no. 6, pp. 838–846, Sep. 2012.

[76] B. D. Gessner, N. Shindo, and S. Briand, “Seasonal influenza epidemiology in sub-Saharan Africa: A systematic review,” *The Lancet Infectious Diseases*, vol. 11, no. 3. Elsevier, pp. 223–235, 01-Mar-2011.

[77] “Human Development Index (HDI) | Human Development Reports.” [Online]. Available: http://hdr.undp.org/en/content/human-development-index-hdi. [Accessed: 03-Jun-2020].

[78] “2020 World Press Freedom Index | RSF.” [Online]. Available: https://rsf.org/en/ranking.

[Accessed: 03-Jun-2020].

[79] R. Amorós, D. Conesa, M. Angel Martinez-Beneito, and A. López-Quílez, “STATISTICAL METHODS FOR DETECTING THE ONSET OF INFLUENZA OUTBREAKS: A REVIEW,” 2015.

[80] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson, and N. Andrews, “Statistical methods for the prospective detection of infectious disease outbreaks: a review,” *J. R. Stat. Soc. Ser. A (Statistics Soc.*, vol. 175, no. 1, pp. 49–82, Jan. 2012.

[81] R. E. Serfling, “Methods for Current Statistical Analysis of Excess Pneumonia-influenza Deaths.”

[82] A. Spreco, O. Eriksson, Ö. Dahlström, and T. Timpka, “Influenza detection and prediction algorithms: comparative accuracy trial in Östergötland county, Sweden, 2008-2012,” 2017.

[83] B. J. Cowling, I. O. L. Wong, L.-M. Ho, S. Riley, and G. M. Leung, “Methods for monitoring influenza surveillance data,” *Int. Epidemiol. Assoc. Int. J. Epidemiol.*, vol. 35, pp. 1314–1321, 2006.

[84] K. M. Neuzil, B. G. Mellen, P. F. Wright, E. F. Mitchel, and M. R. Griffin, “The Effect of Influenza on Hospitalizations, Outpatient Visits, and Courses of Antibiotics in Children,” *N. Engl. J. Med.*, vol. 342, no. 4, pp. 225–231, Jan. 2000.

[85] J. Xing, H. Burkom, and J. Tokars, “Method selection and adaptation for distributed monitoring of infectious diseases for syndromic surveillance,” *J. Biomed. Inform.*, vol. 44, no. 6, pp. 1093–1101, Dec. 2011.

[86] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole, “A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease,” *J. R. Stat. Soc. Ser. A (Statistics Soc.*, vol. 159, no. 3, p. 547, 1996.

[87] H. Zhou, H. Burkom, C. A. Winston, A. Dey, and U. Ajani, “Practical comparison of aberration detection algorithms for biosurveillance systems,” *J. Biomed. Inform.*, vol. 57, pp. 446–455, Oct. 2015.

[88] J. I. Tokars *et al.*, “Enhancing time-series detection algorithms for automated biosurveillance,” *Emerg. Infect. Dis.*, vol. 15, no. 4, pp. 533–539, 2009.

[89] M. Frisén, E. Andersson, and L. Schiöler, “Robust outbreak surveillance of epidemics in Sweden,” *Stat. Med.*, vol. 28, no. 3, pp. 476–493, Feb. 2009.

[90] S. H. Steiner, K. Grant, M. Coory, and H. A. Kelly, “Detecting the start of an influenza outbreak using exponentially weighted moving average charts,” *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, 2010.

[91] L. Hutwagner, W. Thompson, G. M. Seeman, and T. Treadwell, “The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS).”

[92] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, “Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance,” *PLoS Comput. Biol.*, vol. 11, no. 10, 2015.

[93] P. Guo *et al.*, “Developing a dengue forecast model using machine learning: A case study in China,” *PLoS Negl. Trop. Dis.*, vol. 11, no. 10, p. e0005973, Oct. 2017.

[94] N. A. Walton, M. R. Poynton, P. H. Gesteland, C. Maloney, C. Staes, and J. C. Facelli, “Predicting the start week of respiratory syncytial virus outbreaks using real time weather variables,” *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, p. 68, Dec. 2010.

[95] C. Erdman and J. W. Emerson, “A fast Bayesian change point analysis for the segmentation of microarray data,” *Bioinformatics*, vol. 24, no. 19, pp. 2143–2148, Oct. 2008.

[96] D. Barry and J. A. Hartigan, “A Bayesian Analysis for Change Point Problems,” 1993.

[97] T. A. Kass-Hout *et al.*, “Application of change point analysis to daily influenza-like illness emergency department visits,” *J. Am. Med. Informatics Assoc.*, vol. 19, no. 6, pp. 1075–1081, Nov. 2012.

[98] C. Erdman and J. W. Emerson, “bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems Journal of Statistical Software bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems,” 2007.

[99] N. Jafarpour, M. Izadi, D. Precup, and D. L. Buckeridge, “Quantifying the determinants of outbreak detection performance through simulation and machine learning,” *J. Biomed. Inform.*, vol. 53, pp. 180–187, Feb. 2015.

[100] R Core Team, “R: A Language and Environment for Statistical Computing.” Vienna, Austria, 2020.

[101] B. Riplley, B. Venables, D. M. Bates, D. Firth, K. Hornik, and A. Gebhardt, “Package ‘MASS’. Support Functions and Datasets for Venables and Ripley’s MASS.” p. 169, 2018.

[102] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[103] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019.

[104] R. M. O’Brien, “A caution regarding rules of thumb for variance inflation factors,” *Qual. Quant.*, vol. 41, no. 5, pp. 673–690, Oct. 2007.

[105] A. G. Hoen, M. Keller, A. D. Verma, D. L. Buckeridge, and J. S. Brownstein, “Electronic Event–based Surveillance for Monitoring Dengue, Latin America,” *Emerg. Infect. Dis.*, vol. 18, no. 7, pp. 1147–1150, Jul. 2012.

[106] S. Pollett *et al.*, “Clinical Infectious Diseases Evaluating Google Flu Trends in Latin America: Important Lessons for the Next Phase of Digital Disease Detection,” 2016.

[107] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” *Science (80-.).*, vol. 343, no. 6176, pp. 1203–1205, Mar. 2014.

[108] J. R. Ortiz, H. Zhou, D. K. Shay, K. M. Neuzil, A. L. Fowlkes, and C. H. Goss, “Monitoring Influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends,” *PLoS One*, vol. 6, no. 4, 2011.

[109] D. J. McIver and J. S. Brownstein, “Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time,” *PLoS Comput. Biol.*, vol. 10, no. 4, p. e1003581, Apr. 2014.

[110] K. Baltrusaitis *et al.*, “Comparison of crowd-sourced, electronic health records based, and traditional health-care based influenza-tracking systems at multiple spatial resolutions in the United States of America,” *BMC Infect. Dis.*, vol. 18, no. 1, p. 403, Dec. 2018.

[111] R. M. Bloom, D. L. Buckeridge, and K. E. Cheng, “Finding Leading Indicators for Disease Outbreaks: Filtering, Cross-correlation, and Caveats,” *J. Am. Med. Informatics Assoc.*, vol. 14, no. 1, pp. 76–85, Jan. 2007.

[112] G. Shmueli and H. Burkom, “Statistical challenges facing early outbreak detection in biosurveillance,” *Technometrics*, vol. 52, no. 1, pp. 39–51, Feb. 2010.

[113] P. Chakraborty, B. Lewis, S. Eubank, J. S. Brownstein, M. Marathe, and N. Ramakrishnan, “What to know before forecasting the flu,” *PLOS Comput. Biol.*, vol. 14, no. 10, p. e1005964, 2018.

[114] D. Scales, A. Zelenev, and J. S. Brownstein, “Quantifying the effect of media limitations on outbreak data in a global online web-crawling epidemic intelligence system, 2008-2011.,” *Emerg. Health Threats J.*, vol. 6, p. 21621, 2013.

7 Appendix

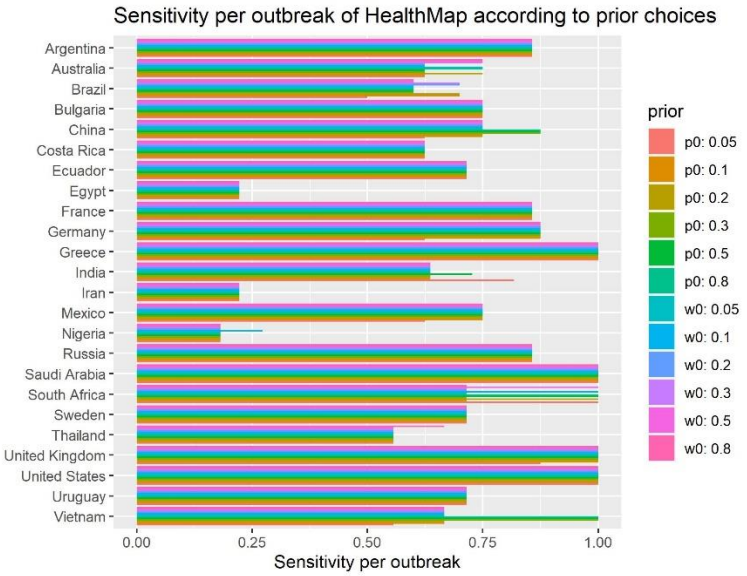

Figure 9: Testing of bcp priors for HealthMap.

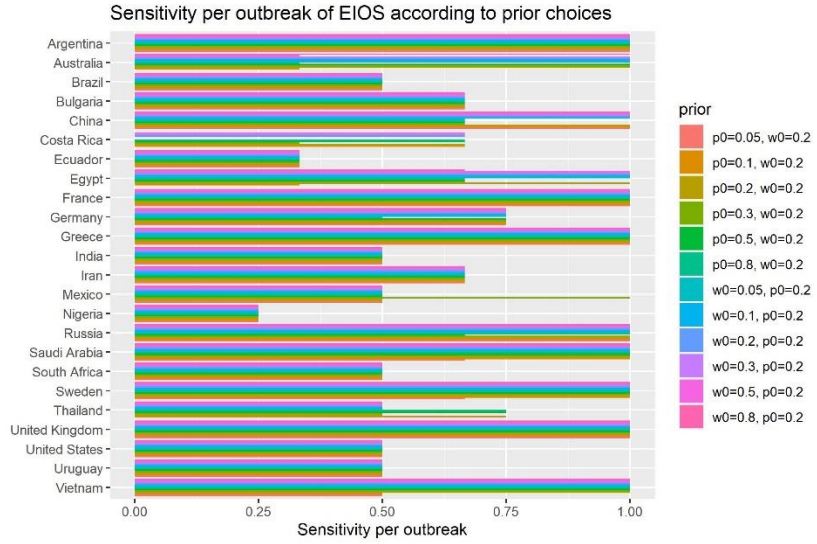

Figure 10: Testing of bcp priors for EIOS.

	HealthMap best cutoff		EIOS best cutoff	
country	**sensitivity + FAR**	**timeliness + FAR**	**sensitivity + FAR**	**timeliness + FAR**
Argentina	0.1	0.45 - 0.6	0.25	0.6 - 0.7
Australia	0.1	0.15	0.2	0.2
Brazil	0.1	0.1	0.3 - 0.35	0.3 - 0.65
Bulgaria	0.1 - 0.9	0.1 - 0.9	0.2 - 0.25	0.45 - 0.9
China	0.3	0.7 - 0.75	0.35 - 0.4	0.45 - 0.7
Costa Rica	0.1 - 0.9	0.1 - 0.9	0.2	0.2
Ecuador	0.1 - 0.9	0.1 - 0.9	0.45 - 0.8	0.45 - 0.9
Egypt	0.1 - 0.45	0.6 - 0.75	0.15 - 0.9	0.15
France	0.9	0.15	0.1	0.45 - 0.65
Germany	0.1 - 0.9	0.1 - 0.9	0.3	0.1
Greece	0.1 - 0.9	0.1 - 0.9	0.75 - 0.9	0.75 - 0.9
India	0.55	0.8	0.75 - 0.8	0.75 - 0.8
Iran	0.1	0.1	0.1	0.1
Mexico	0.3	0.3	0.25	0.1
Nigeria	0.1 - 0.9	0.1 - 0.9	0.15	0.1
Russia	0.1	0.3 - 0.35	0.2 - 0.25	0.2 - 0.3
Saudi Arabia	0.1 - 0.9	0.1 - 0.9	0.75 - 0.9	0.15 - 0.45
South Africa	0.1 - 0.9	0.1 - 0.9	0.1	0.15
Sweden	0.1 - 0.9	0.1 - 0.9	0.45 - 0.65	0.45 - 0.65
Thailand	0.1 - 0.9	0.1 - 0.9	0.2 - 0.25	0.2 - 0.25
United Kingdom	0.2	0.25 - 0.5	0.25 - 0.5	0.1 - 0.65
United States	0.5	0.5	0.15	0.8 - 0.9
Uruguay	0.1 - 0.9	0.1 - 0.9	0.15	0.15
Vietnam	0.35 - 0.4	0.9	0.15 - 0.2	0.15 - 0.2
Average cutoff	**0.48**	**0.50**	**0.45**	**0.49**

Table 7: Best cutoffs of posterior probability for outbreak detection in HealthMap and EIOS. *The best cutoff point was determined by plotting ROC curves and determining the point with the least Euclidean distance from the optimal point (sensitivity/timeliness = 100%, false alarm rate (FAR) = 0%).*

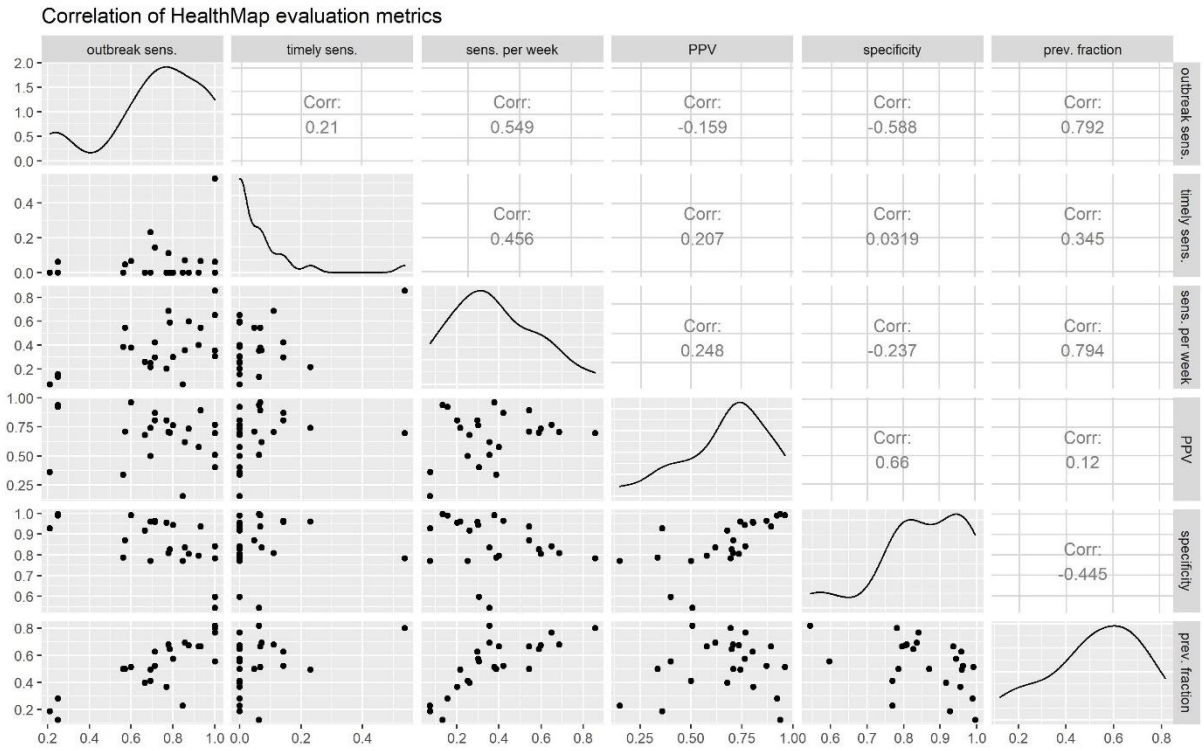

Figure 11: Correlation of HealthMap evaluation metrics.

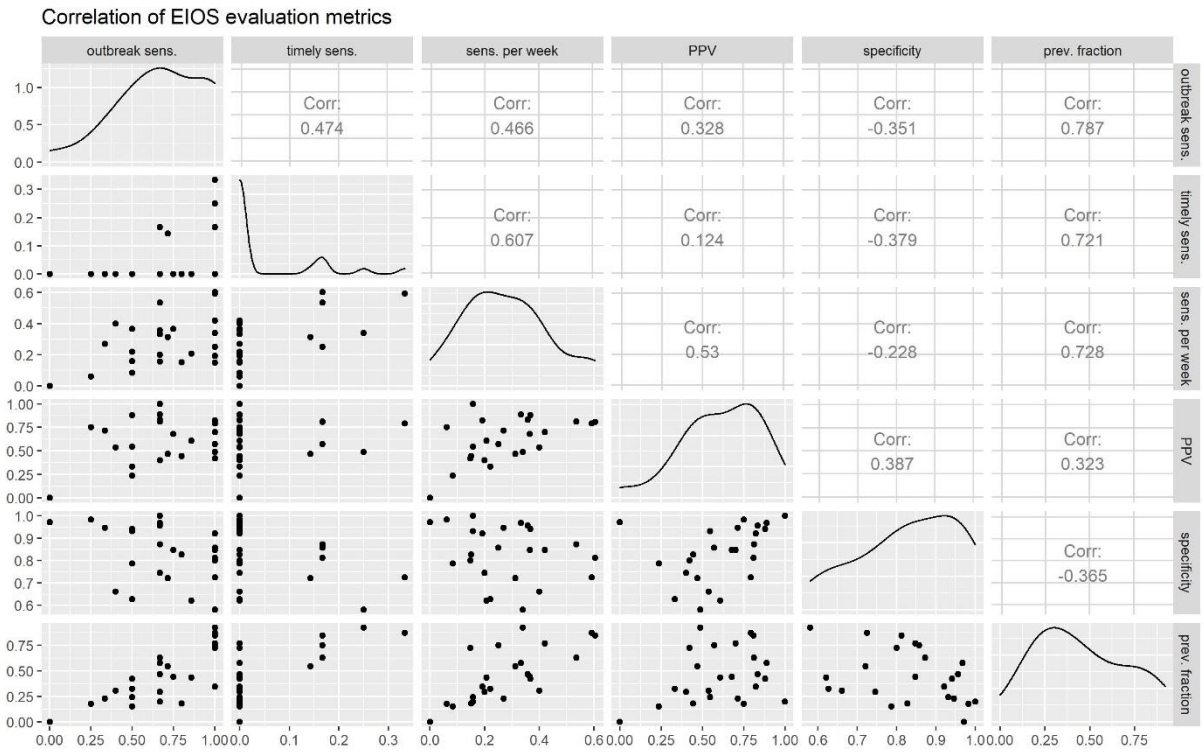

Figure 12: Correlation of EIOS evaluation metrics.

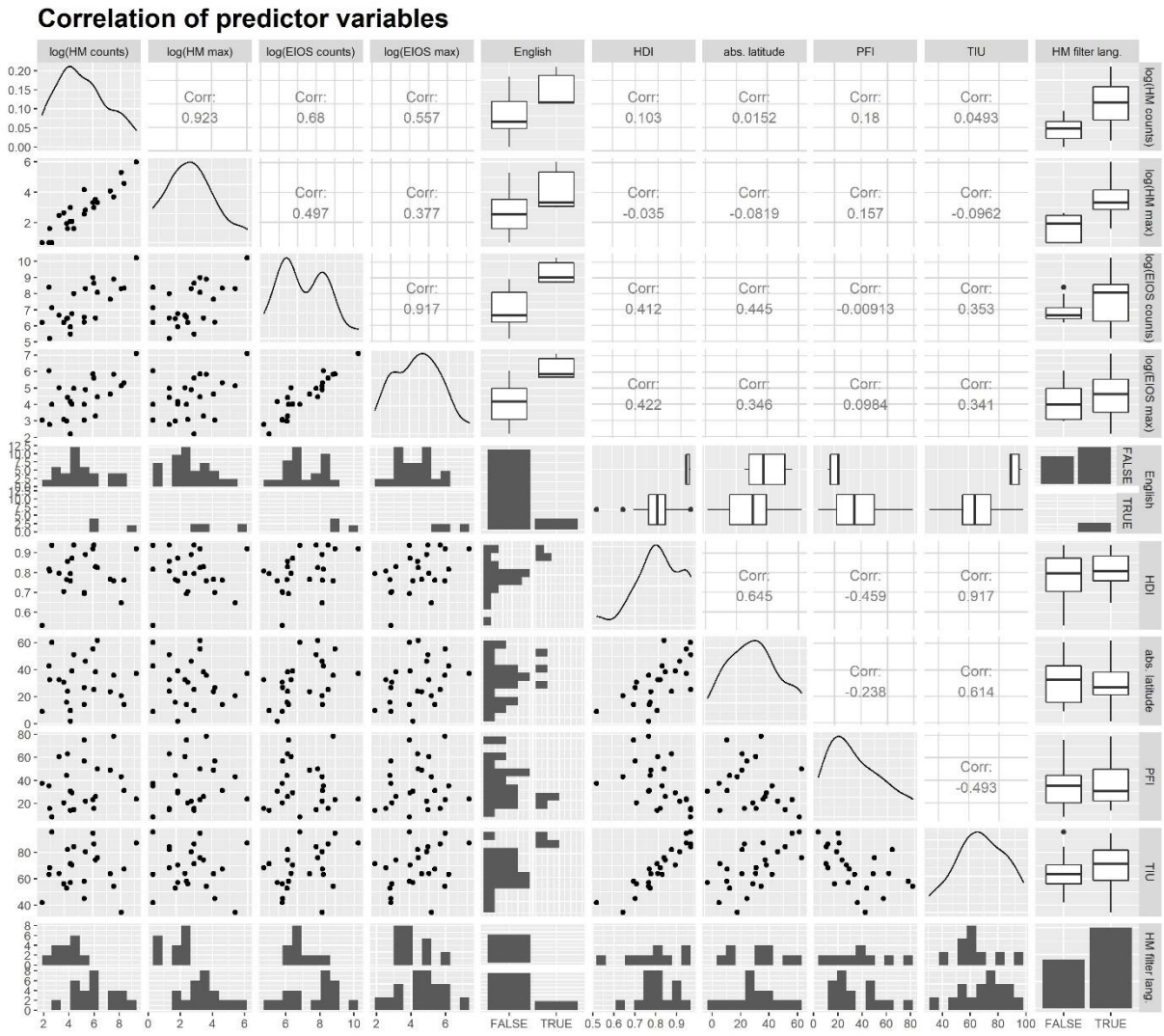

Figure 13: Pairwise correlation of predictor variables. *In case of categorical predictors, boxplots and faceted histograms were plotted, in case of continuous predictors, scatterplots were drawn.*

Table 8: 95% confidence intervals of all simple evaluation metrics.

Country	EBS system	Sensitivity per outbreak	Timely sensitivity	Sensitivity per week	Positive predictive value	Specificity	Prevented fraction
Argentina	HealthMap	0.753 - 1	0 - 0.247	0.57 - 0.727	0.685 - 0.837	0.781 - 0.89	0.708 - 0.819
Argentina	EIOS	0.398 - 1	0 - 0.602	0.282 - 0.568	0.506 - 0.853	0.73 - 0.928	0.648 - 0.856
Australia	HealthMap	0.419 - 0.916	0.018 - 0.428	0.22 - 0.387	0.661 - 0.906	0.923 - 0.981	0.54 - 0.706
Australia	EIOS	0.284 - 0.995	0 - 0.522	0.067 - 0.276	0.215 - 0.692	0.666 - 0.888	0.124 - 0.252
Brazil	HealthMap	0.524 - 0.936	0.014 - 0.347	0.601 - 0.763	0.622 - 0.782	0.748 - 0.86	0.608 - 0.741
Brazil	EIOS	0.223 - 0.957	0.004 - 0.641	0.397 - 0.67	0.648 - 0.92	0.712 - 0.922	0.494 - 0.745
Bulgaria	HealthMap	0.519 - 0.957	0 - 0.218	0.225 - 0.389	0.625 - 0.872	0.903 - 0.97	0.488 - 0.655
Bulgaria	EIOS	0.223 - 0.957	0 - 0.459	0.216 - 0.52	0.586 - 0.964	0.878 - 0.991	0.333 - 0.607
China	HealthMap	0.617 - 0.984	0 - 0.206	0.519 - 0.675	0.651 - 0.808	0.739 - 0.86	0.599 - 0.741
China	EIOS	0.541 - 1	0.004 - 0.641	0.136 - 0.396	0.34 - 0.782	0.746 - 0.933	0.648 - 0.83
Costa Rica	HealthMap	0.384 - 0.882	0 - 0.218	0.19 - 0.342	0.537 - 0.801	0.869 - 0.95	0.32 - 0.482

Country	EBS system	Sensitivity per outbreak	Timely sensitivity	Sensitivity per week	Positive predictive value	Specificity	Prevented fraction
Costa Rica	EIOS	0 - 0.522	0 - 0.522	0 - 0.088	0 - 0.842	0.88 - 0.991	0.025 - 0.025
Ecuador	HealthMap	0.419 - 0.916	0.018 - 0.428	0.335 - 0.512	0.761 - 0.943	0.927 - 0.984	0.431 - 0.613
Ecuador	EIOS	0.043 - 0.777	0 - 0.459	0.138 - 0.441	0.419 - 0.916	0.867 - 0.985	0.155 - 0.322
Egypt	HealthMap	0.073 - 0.524	0 - 0.206	0.103 - 0.224	0.749 - 0.991	0.962 - 0.999	0.223 - 0.349
Egypt	EIOS	0.223 - 0.957	0 - 0.459	0.1 - 0.337	0.211 - 0.613	0.616 - 0.85	0.199 - 0.41
France	HealthMap	0.546 - 0.981	0 - 0.247	0.034 - 0.133	0.072 - 0.27	0.708 - 0.824	0.178 - 0.287
France	EIOS	0.398 - 1	0.006 - 0.806	0.221 - 0.474	0.329 - 0.649	0.432 - 0.718	0.889 - 0.944
Germany	HealthMap	0.492 - 0.953	0 - 0.232	0.503 - 0.673	0.609 - 0.782	0.767 - 0.876	0.563 - 0.723
Germany	EIOS	0.29 - 0.963	0.004 - 0.579	0.187 - 0.463	0.291 - 0.653	0.592 - 0.829	0.421 - 0.659
Greece	HealthMap	0.572 - 0.982	0.002 - 0.339	0.279 - 0.44	0.507 - 0.723	0.776 - 0.885	0.617 - 0.762
Greece	EIOS	0.398 - 1	0 - 0.602	0.066 - 0.271	0.203 - 0.665	0.67 - 0.896	0.583 - 0.829
India	HealthMap	0.34 - 0.782	0.001 - 0.238	0.453 - 0.633	0.607 - 0.797	0.818 - 0.912	0.427 - 0.573
India	EIOS	0.157 - 0.843	0 - 0.369	0.246 - 0.501	0.688 - 0.975	0.786 - 0.967	0.309 - 0.541
Iran	HealthMap	0.073 - 0.524	0.002 - 0.302	0.077 - 0.211	0.698 - 0.998	0.976 - 1	0.101 - 0.151
Iran	EIOS	0.223 - 0.957	0 - 0.459	0.07 - 0.286	0.631 - 1	0.94 - 1	0.136 - 0.279
Mexico	HealthMap	0.323 - 0.837	0.002 - 0.319	0.296 - 0.467	0.868 - 0.995	0.966 - 0.999	0.427 - 0.598
Mexico	EIOS	0.349 - 0.968	0 - 0.369	0.236 - 0.51	0.476 - 0.841	0.73 - 0.928	0.327 - 0.562
Nigeria	HealthMap	0.061 - 0.456	0 - 0.176	0.034 - 0.135	0.18 - 0.575	0.884 - 0.958	0.153 - 0.226
Nigeria	EIOS	0.032 - 0.651	0 - 0.369	0.013 - 0.169	0.194 - 0.994	0.887 - 0.996	0.127 - 0.234
Russia	HealthMap	0.64 - 0.998	0 - 0.247	0.318 - 0.486	0.473 - 0.677	0.734 - 0.849	0.584 - 0.738
Russia	EIOS	0.223 - 0.957	0 - 0.459	0.204 - 0.484	0.653 - 0.986	0.89 - 0.996	0.438 - 0.705
Saudi Arabia	HealthMap	0.541 - 1	0 - 0.459	0.236 - 0.384	0.314 - 0.494	0.521 - 0.669	0.416 - 0.686
Saudi Arabia	EIOS	0.421 - 0.996	0 - 0.41	0.126 - 0.311	0.406 - 0.785	0.423 - 0.793	0.32 - 0.554
South Africa	HealthMap	0.462 - 0.95	0 - 0.247	0.144 - 0.272	0.651 - 0.912	0.913 - 0.98	0.287 - 0.452
South Africa	EIOS	0.068 - 0.932	0 - 0.602	0.023 - 0.2	0.068 - 0.499	0.663 - 0.881	0.097 - 0.224
Sweden	HealthMap	0.386 - 0.909	0 - 0.247	0.187 - 0.325	0.387 - 0.613	0.701 - 0.829	0.327 - 0.502
Sweden	EIOS	0.541 - 1	0.004 - 0.641	0.472 - 0.724	0.667 - 0.909	0.674 - 0.911	0.784 - 0.892
Thailand	HealthMap	0.299 - 0.802	0 - 0.206	0.276 - 0.506	0.239 - 0.447	0.731 - 0.833	0.413 - 0.584
Thailand	EIOS	0.157 - 0.843	0 - 0.369	0.06 - 0.313	0.234 - 0.833	0.812 - 0.961	0.174 - 0.328
United Kingdom	HealthMap	0.681 - 0.998	0.002 - 0.319	0.466 - 0.621	0.817 - 0.945	0.888 - 0.968	0.589 - 0.734
United Kingdom	EIOS	0.541 - 1	0.043 - 0.777	0.468 - 0.707	0.659 - 0.892	0.561 - 0.854	0.822 - 0.908
United States	HealthMap	0.753 - 1	0.251 - 0.808	0.782 - 0.912	0.616 - 0.766	0.721 - 0.836	0.749 - 0.844
United States	EIOS	0.053 - 0.853	0 - 0.522	0.27 - 0.541	0.374 - 0.693	0.522 - 0.782	0.206 - 0.424
Uruguay	HealthMap	0.386 - 0.909	0.05 - 0.538	0.147 - 0.301	0.567 - 0.875	0.924 - 0.981	0.402 - 0.585
Uruguay	EIOS	0.068 - 0.932	0 - 0.602	0.115 - 0.36	0.18 - 0.518	0.491 - 0.75	0.2 - 0.478
Vietnam	HealthMap	0.794 - 1	0.002 - 0.302	0.288 - 0.427	0.42 - 0.594	0.46 - 0.626	0.773 - 0.853
Vietnam	EIOS	0.478 - 1	0 - 0.522	0.109 - 0.301	0.566 - 0.962	0.786 - 0.983	0.235 - 0.474

Table 9: Regression coefficients from univariable regressions for HealthMap. *Predictors with p < 0.2 are highlighted in orange.*

outcome	predictor	Coefficient	p-value	95% confidence interval	R^2 value	diagnostic criteria violated
sensitivity per outbreak	total counts (categorical, linear effect)	0.059	0.4549	-0.102 - 0.2194	0.041	
	total counts (categorical, quadratic effect)	-0.033	0.6958	-0.2091 - 0.1423	0.041	
	total counts	0.028	0.2710	-0.0238 - 0.0805	0.057	
	maximum weekly counts	-0.002	0.9617	-0.0722 - 0.0689	0.000	influential outlier
	global region: temp.Southern	0.007	0.9533	-0.2417 - 0.2559	0.094	
	global region: tropical	-0.141	0.1864	-0.3556 - 0.0739	0.094	
	english: TRUE	0.151	0.2527	-0.1159 - 0.4176	0.062	
	HDI.2018	0.889	0.0870	-0.1406 - 1.9177	0.133	
	latitude	0.003	0.2287	-0.0023 - 0.0091	0.068	
	longitude	0.000	0.7748	-0.0015 - 0.0011	0.004	
	PFI.2018	-0.001	0.6638	-0.0056 - 0.0037	0.009	
	total internet users	0.006	0.0211	0.0011 - 0.0117	0.228	
	HM filter language: TRUE	0.056	0.5503	-0.1367 - 0.2494	0.017	
timely sensitivity	total counts (categorical, linear effect)	0.060	0.1623	-0.0262 - 0.1464	0.099	Normality
	total counts (categorical, quadratic effect)	0.032	0.4902	-0.0632 - 0.1276	0.099	Normality
	total counts	0.026	0.0469	0.0004 - 0.0524	0.168	Normality
	maximum weekly counts	0.038	0.0281	0.0045 - 0.0717	0.201	Normality
	global region: temp.Southern	0.036	0.6114	-0.1099 - 0.1823	0.013	Normality
	global region: tropical	0.004	0.9484	-0.116 - 0.1236	0.013	Normality
	english: TRUE	0.211	0.0018	0.0882 - 0.3347	0.365	Normality
	HDI.2018	0.299	0.2267	-0.1994 - 0.7968	0.066	Normality
	latitude	-0.001	0.7357	-0.0037 - 0.0027	0.005	Normality
	longitude	-0.001	0.0564	-0.0013 - 0	0.156	Normality
	PFI.2018	-0.001	0.3009	-0.0039 - 0.0013	0.049	Normality
	total internet users	0.002	0.2481	-0.0013 - 0.0049	0.060	Normality
	HM filter language: TRUE	0.068	0.1779	-0.0333 - 0.1693	0.081	Normality
sensitivity per week	total counts (categorical, linear effect)	0.202	0.0031	0.076 - 0.3277	0.351	
	total counts (categorical, quadratic effect)	0.072	0.2948	-0.0673 - 0.211	0.351	
	total counts	0.082	0.0000	0.0488 - 0.1149	0.545	
	maximum weekly counts	0.086	0.0026	0.0333 - 0.1381	0.344	influential outlier
	global region: temp.Southern	-0.048	0.6949	-0.2999 - 0.2036	0.008	
	global region: tropical	-0.019	0.8530	-0.225 - 0.1878	0.008	
	english: TRUE	0.216	0.0840	-0.0315 - 0.4644	0.130	
	HDI.2018	0.533	0.2091	-0.321 - 1.3864	0.071	
	latitude	0.001	0.6176	-0.0041 - 0.0068	0.012	
	longitude	-0.001	0.3120	-0.0019 - 0.0006	0.046	
	PFI.2018	0.000	0.9339	-0.0047 - 0.0043	0.000	
	total internet users	0.003	0.3224	-0.0028 - 0.008	0.045	
	HM filter language: TRUE	0.131	0.1272	-0.0405 - 0.3035	0.103	
positive predictive value	total counts (categorical, linear effect)	0.062	0.4235	-0.0956 - 0.2191	0.031	
	total counts (categorical, quadratic effect)	0.010	0.9059	-0.164 - 0.184	0.031	
	total counts	0.025	0.2971	-0.0237 - 0.0741	0.049	
	maximum weekly counts	0.051	0.1003	-0.0107 - 0.1134	0.118	

outcome	predictor	Coefficient	p-value	95% confidence interval	R^2 value	diagnostic criteria violated
	global region: temp.Southern	0.118	0.3383	-0.1326 - 0.3689	0.059	Normality
	global region: tropical	-0.029	0.7709	-0.2347 - 0.1764	0.059	Normality
	english: TRUE	0.142	0.2765	-0.1222 - 0.4068	0.054	
	HDI.2018	0.044	0.9207	-0.8617 - 0.9496	0.000	influential outlier
	latitude	0.000	0.9132	-0.0059 - 0.0053	0.001	
	longitude	-0.001	0.2161	-0.002 - 0.0005	0.069	
	PFI.2018	0.000	0.9930	-0.0046 - 0.0046	0.000	
	total internet users	-0.001	0.7433	-0.0065 - 0.0047	0.005	
	HM filter language: TRUE	0.093	0.2968	-0.0878 - 0.2744	0.049	
specificity	total counts (categorical, linear effect)	-0.022	0.6207	-0.1117 - 0.0682	0.019	homoskedasticity
	total counts (categorical, quadratic effect)	0.014	0.7665	-0.085 - 0.1138	0.019	homoskedasticity
	total counts	-0.010	0.4567	-0.0384 - 0.0178	0.025	homoskedasticity
	maximum weekly counts	0.003	0.8686	-0.0345 - 0.0406	0.001	homoskedasticity
	global region: temp.Southern	0.086	0.2204	-0.0554 - 0.2268	0.077	
	global region: tropical	-0.001	0.9883	-0.1165 - 0.1149	0.077	
	english: TRUE	0.041	0.5895	-0.1129 - 0.1939	0.013	homoskedasticity
	HDI.2018	-0.163	0.5152	-0.6721 - 0.347	0.020	homoskedasticity
	latitude	-0.001	0.7400	-0.0037 - 0.0027	0.005	homoskedasticity
	longitude	-0.001	0.1474	-0.0012 - 2e-04	0.093	homoskedasticity
	PFI.2018	-0.002	0.1162	-0.0044 - 5e-04	0.108	homoskedasticity
	total internet users	-0.002	0.3014	-0.0047 - 0.0015	0.048	
	HM filter language: TRUE	0.023	0.6558	-0.0821 - 0.1279	0.009	homoskedasticity
prevented fraction	total counts (categorical, linear effect)	0.112	0.0916	-0.0198 - 0.2435	0.140	Normality
	total counts (categorical, quadratic effect)	-0.004	0.9526	-0.1482 - 0.1399	0.140	Normality
	total counts	0.047	0.0286	0.0054 - 0.0881	0.208	
	maximum weekly counts	0.028	0.3384	-0.0316 - 0.0879	0.044	
	global region: temp.Southern	0.017	0.8791	-0.2089 - 0.2422	0.006	
	global region: tropical	-0.023	0.8041	-0.2182 - 0.1712	0.006	
	english: TRUE	0.177	0.1154	-0.0472 - 0.4016	0.114	
	HDI.2018	0.378	0.4133	-0.5634 - 1.319	0.032	
	latitude	0.001	0.7940	-0.0045 - 0.0058	0.003	
	longitude	0.000	0.9953	-0.0011 - 0.0011	0.000	
	PFI.2018	0.000	0.8238	-0.0036 - 0.0045	0.002	
	total internet users	0.003	0.3060	-0.0025 - 0.0077	0.050	
	HM filter language: TRUE	0.042	0.6104	-0.1258 - 0.2091	0.013	

Table 10: Regression coefficients from univariable regressions for EIOS. *Predictors with $p < 0.2$ are highlighted in orange.*

outcome	predictor	Coefficient	p-value	95% confidence interval	R^2 value	diagnostic criteria violated
sensitivity per outbreak	total counts (categorical, linear effect)	0.106	0.2905	-0.097 - 0.3083	0.086	
	total counts (categorical, quadratic effect)	-0.074	0.4980	-0.2983 - 0.1497	0.086	
	total counts	0.069	0.1273	-0.0213 - 0.1596	0.102	influential outlier
	maximum weekly counts	0.074	0.1141	-0.0194 - 0.1681	0.110	influential outlier
	global region: temp.Southern	-0.048	0.7637	-0.3742 - 0.2786	0.093	
	global region: tropical	-0.188	0.1582	-0.4559 - 0.0793	0.093	
	english: TRUE	0.055	0.7527	-0.3043 - 0.4149	0.005	

outcome	predictor	Coefficient	p-value	95% confidence interval	R^2 value	diagnostic criteria violated
	HDI.2018	1.165	0.0364	0.0806 - 2.2502	0.184	
	latitude	0.009	0.0048	0.0032 - 0.0155	0.309	
	longitude	0.001	0.0979	-3e-04 - 0.003	0.120	
	PFI.2018	0.003	0.2904	-0.0029 - 0.0091	0.051	homoskedasticity
	total internet users	0.007	0.0613	-3e-04 - 0.0135	0.150	
timely sensitivity	total counts (categorical, linear effect)	0.079	0.0160	0.0162 - 0.1409	0.296	Normality
	total counts (categorical, quadratic effect)	-0.031	0.3638	-0.0997 - 0.0382	0.296	Normality
	total counts	0.035	0.0239	0.0051 - 0.0646	0.211	Normality
	maximum weekly counts	0.029	0.0776	-0.0035 - 0.0613	0.135	Normality
	global region: temp.Southern	-0.082	0.1458	-0.1937 - 0.0307	0.129	Normality
	global region: tropical	-0.058	0.2063	-0.1497 - 0.0343	0.129	Normality
	english: TRUE	0.069	0.2590	-0.0542 - 0.1914	0.058	Normality
	HDI.2018	0.382	0.0521	-0.0038 - 0.7681	0.161	Normality
	latitude	0.003	0.0177	5e-04 - 0.0051	0.230	Normality
	longitude	0.000	0.8947	-6e-04 - 6e-04	0.001	Normality
	PFI.2018	-0.001	0.3188	-0.0031 - 0.0011	0.045	Normality
	total internet users	0.003	0.0235	4e-04 - 0.0051	0.212	Normality
sensitivity per week	total counts (categorical, linear effect)	0.145	0.0035	0.0533 - 0.2359	0.443	
	total counts (categorical, quadratic effect)	-0.088	0.0836	-0.1892 - 0.0128	0.443	
	total counts	0.070	0.0041	0.0248 - 0.1159	0.318	
	maximum weekly counts	0.068	0.0086	0.0191 - 0.1168	0.274	
	global region: temp.Southern	-0.097	0.3016	-0.2887 - 0.0939	0.067	
	global region: tropical	-0.067	0.3820	-0.2242 - 0.0895	0.067	
	english: TRUE	0.115	0.2496	-0.0868 - 0.317	0.060	
	HDI.2018	0.664	0.0394	0.0353 - 1.2925	0.179	
	latitude	0.005	0.0081	0.0015 - 0.0088	0.278	
	longitude	-0.001	0.2821	-0.0015 - 5e-04	0.052	
	PFI.2018	-0.002	0.3406	-0.0051 - 0.0018	0.041	homoskedasticity
	total internet users	0.004	0.0246	6e-04 - 0.0083	0.209	
positive predictive value	total counts (categorical, linear effect)	0.085	0.3194	-0.0887 - 0.2593	0.093	
	total counts (categorical, quadratic effect)	-0.080	0.3987	-0.272 - 0.1127	0.093	
	total counts	0.059	0.1331	-0.0194 - 0.1368	0.100	Normality
	maximum weekly counts	0.085	0.0315	0.0083 - 0.1621	0.194	Normality
	global region: temp.Southern	-0.148	0.2305	-0.3978 - 0.1014	0.286	
	global region: tropical	0.213	0.0424	0.008 - 0.4173	0.286	
	english: TRUE	-0.026	0.8643	-0.3364 - 0.2846	0.001	
	HDI.2018	-0.293	0.5597	-1.3205 - 0.7338	0.016	
	latitude	0.002	0.4393	-0.0039 - 0.0087	0.027	
	longitude	0.001	0.2606	-6e-04 - 0.0023	0.057	
	PFI.2018	0.005	0.0543	-1e-04 - 0.0096	0.158	
	total internet users	-0.001	0.7533	-0.0074 - 0.0055	0.005	
specificity	total counts (categorical, linear effect)	-0.023	0.6304	-0.1184 - 0.0734	0.020	
	total counts (categorical, quadratic effect)	0.017	0.7429	-0.0891 - 0.123	0.020	
	total counts	-0.018	0.3982	-0.0607 - 0.0251	0.033	
	maximum weekly counts	-0.016	0.4756	-0.0605 - 0.0292	0.023	
	global region: temp.Southern	-0.017	0.7783	-0.145 - 0.1101	0.337	
	global region: tropical	0.153	0.0063	0.0479 - 0.2571	0.337	
	english: TRUE	-0.107	0.1739	-0.2647 - 0.0509	0.082	

outcome	predictor	Coefficient	p-value	95% confidence interval	R^2 value	diagnostic criteria violated
	HDI.2018	-0.641	0.0097	-1.1112 - -0.1715	0.267	
	latitude	-0.003	0.0778	-0.006 - 3e-04	0.135	
	longitude	0.000	0.2959	-4e-04 - 0.0012	0.050	
	PFI.2018	0.002	0.1844	-9e-04 - 0.0045	0.079	
	total internet users	-0.004	0.0127	-0.0068 - -9e-04	0.251	
prevented fraction	total counts (categorical, linear effect)	0.149	0.0844	-0.022 - 0.3193	0.270	
	total counts (categorical, quadratic effect)	-0.160	0.0918	-0.3489 - 0.0284	0.270	
	total counts	0.089	0.0319	0.0085 - 0.1701	0.193	Normality
	maximum weekly counts	0.075	0.0909	-0.0129 - 0.1622	0.124	Normality
	global region: temp.Southern	-0.196	0.1757	-0.4875 - 0.095	0.187	
	global region: tropical	-0.232	0.0563	-0.4708 - 0.0068	0.187	
	english: TRUE	0.001	0.9970	-0.339 - 0.3402	0.000	
	HDI.2018	1.143	0.0287	0.1302 - 2.1552	0.199	
	latitude	0.011	0.0003	0.0056 - 0.0159	0.459	
	longitude	0.000	0.9399	-0.0016 - 0.0017	0.000	
	PFI.2018	-0.001	0.7792	-0.0066 - 0.005	0.004	homoskedasticity
	total internet users	0.007	0.0337	6e-04 - 0.0133	0.189	

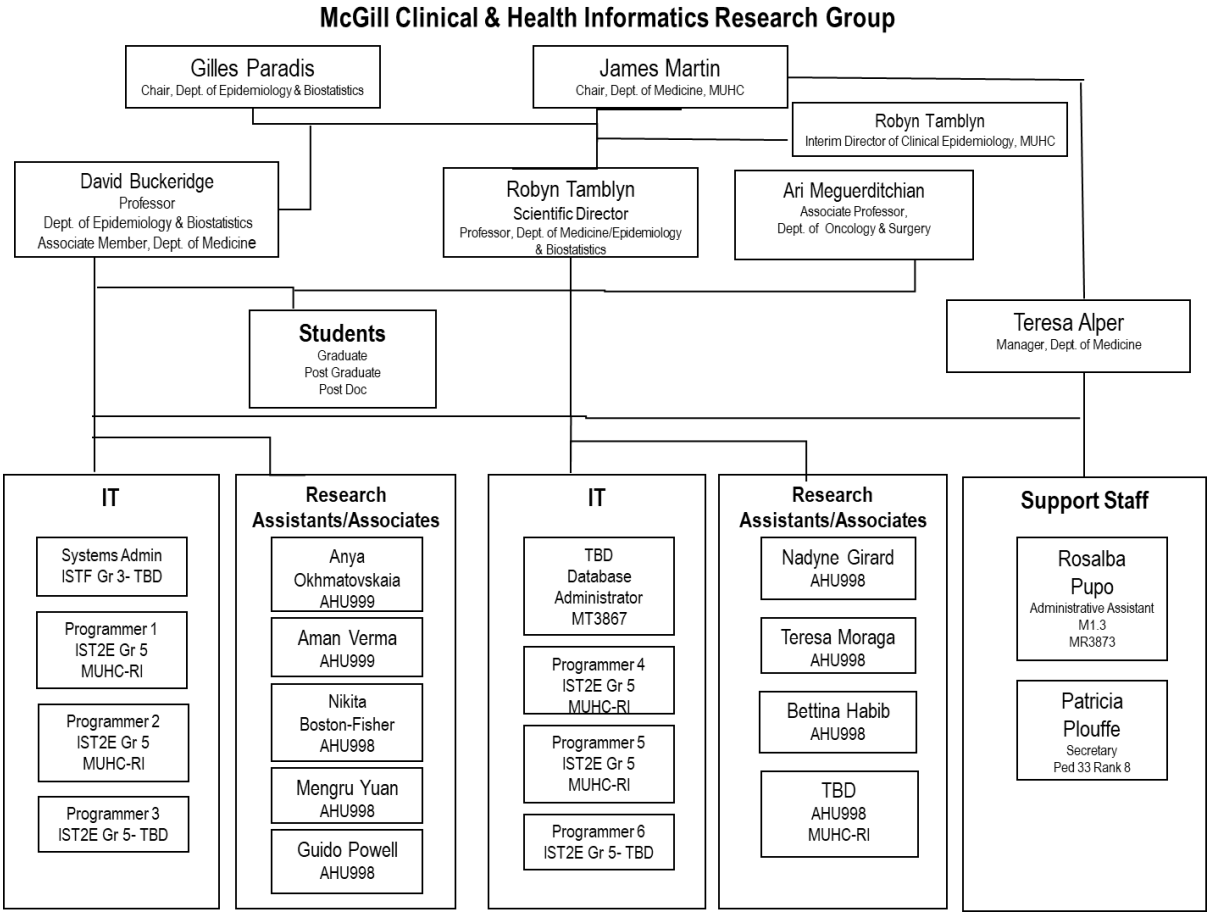

Figure 14: Organisational chart of the McGill Clinical & Health Informatics Research group