Distribution Agreement

Signature:

________________________________ ________________
Roxanne Moore Date

Variability in case and mortality between WHO/MOH and HealthMap curated news reports during the West Africa Ebola outbreak (March 14 - August 28, 2014)

By

Roxanne Moore
Master in Public Health

Dr. Eli Rosenberg
Faculty Thesis Advisor

Variability in case and mortality between WHO/MOH and HealthMap curated news reports during the West Africa Ebola outbreak (March 14 - August 28, 2014)

By

Roxanne Moore
B.A., Christopher Newport University, 2011

Faculty Thesis Advisor: Eli Rosenberg, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2015

Abstract

Variability in case and mortality between WHO/MOH and HealthMap curated news reports during the West Africa Ebola outbreak (March 14 - August 28, 2014)

By Roxanne Moore

Background: The public is more likely to obtain information regarding Ebola through the news as compared to scientific journals or official reports. The manner in which news media portrays scientific information, particularly case and mortality values has an impact upon trust in formal health agencies, public donations, and fear mongering. Previous research has demonstrated that HealthMap curated news articles can effectively be used as a sentinel surveillance system for early detection of infectious diseases (1-6).

Methods: This study uses HealthMap identified news articles as a representative sample of online news to evaluate the variation between news-derived case and mortality as compared to the World Health Organization (WHO) and Ministry of Health (MOH) official reports via the Humanitarian Data Exchange (HDX) sub-national time series dataset.

Results: HealthMap counts, country, and reoriented date accurately predicts 75.9% of the change in estimate for WHO/MOH cases and 90.3% for deaths. When limited to news articles providing a citation for case and mortality counts, prediction increases to 92.2% for cases and 95.9% for deaths with no statistically significant difference in news-derived and WHO/MOH official estimates for cases under the subset model.

Discussion: It is hypothesized that lower predictive capacity for cases is related to greater variability in case estimates due to multiple definitions (suspected, probable, and confirmed cases) as well as changes in case reporting for Sierra Leone during the time of study. Two strengths of this study included evaluating a secondary use of HealthMap, and quantifying distrust in media reported values. Limitations include non-repetitive news article sources and non-longitudinal analysis, as well as inaccuracies in the official source. As a result, differences in news and official counts may be caused by the official reports rather than news.

Future Directions: Five alternative analyses are highlighted: case and death comparisons, rumors, regional-level reporting, longitudinal time series, and incidence analysis. News and official variability analysis may not directly facilitate health responders, but instead evaluate public perception. News-derived predictive models should not be used as a supplement to official estimates. Rather, news articles should link the public to official reports; therefore reducing variability in the public domain.

Variability in case and mortality between WHO/MOH, and HealthMap curated news reports during the West Africa Ebola outbreak (March 14 - August 28, 2014)

By

Roxanne Moore
B.A., Christopher Newport University, 2011

Faculty Thesis Advisor: Eli Rosenberg, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2015

Table of Contents

Background/Literature Review

A. Introduction

The 2014-2015 Ebola Virus Disease (EVD) outbreak in West Africa was the first of its kind within the region, and the largest outbreak of EVD to date (7). The first case of EVD in West Africa has been retrospectively traced to December 2, 2013 (8). However, the disease was not publicly identified until over three months later on March 14, 2014 through a Guinean newspaper and Guinean Ministry of Health report (9). Nine days following on March 23, 2014 the first World Health Organization (WHO) report on EVD was released (10). On March 23, WHO reported 49 cases of the disease including 29 deaths in Guinea, resulting in a Case Fatality Ratio (CFR) of 59% (10). On August 8, 2014 the WHO declared the epidemic a "public health emergency of international concern" (11). By August 28, 2014 three months after WHO formally announced EVD within Sierra Leone (12), 3,052 cases and 1,546 deaths were reported across West Africa (13). The total of 3,052 cases (including 1,739 confirmed, 852 probable, and 461 suspected cases respectively) as well as 1,546 deaths (concerning 892 confirmed, 476 probable, and 178 suspected deaths respectively) were recorded in Guinea, Liberia, and Sierra Leone (excluding Nigeria) (13). The CFR across the three countries was 51%, but this varied from 42% in Sierra Leone, to 50% in Liberia, and 66% in Guinea (13). This section examines the current literature on EVD, how this outbreak has differed from historical EVD outbreaks, and how online news surveillance has played a part in identifying and tracking the disease.

B. Ebola Virus Disease (EVD)

1. EVD transmission

On December 6, 2013, a two-year boy in the remote region of Meliandou, Guinea died of a mysterious illness characterized by fever, black stools, and vomiting later identified as the *Zaire ebolavirus*, commonly referred to as Ebola Virus Disease (EVD) (8). It was hypothesized that the boy contracted the disease while playing in a hollow tree housing a colony of fruit bats, thereby becoming exposed through the bat excrement (14). Although the initial case was not attributable to bush meat consumption, it is plausible that fruit bat bush meat consumption contributed to the diseases further propagation (14, 15). As a result, it is further suspected that the virus was transmitted for months before the outbreak gained recognition in March due to a cluster of cases in the hospitals of Gueckedou and Macenta, Guinea (8, 16).

Although fruit bats served as an intermediate vector, the human transmission chain is driven via direct contact with infected persons' body fluids. Ebola virus RNA can be found in blood, saliva, feces, tears, ocular fluid, vaginal fluid, semen, breast milk, and on the skin (17). Following disease onset, EVD may be capable of persisting in survivors' mucosal membranes for at least 33 days (18). Ebola virus RNA has been detected up to 199 days after symptom onset in semen (19), 78 days in ocular fluid (20), 33 days in vaginal fluid (18), 29 days in stool (18), 22 days in tears (18), 15 days in breast milk (21), 8 days in saliva (21), and 6 days on skin (21). Ebola virus was not detected in samples of vomit, sputum, sweat, urine, or from body louse (18, 21, 22). Given the relatively small number of patients studied, infection durations may not be representative (18). This is of special concern for women breastfeeding as only one mother was examined (21). As EVD is spread through contact with body fluids, caregiving of the ill and burial ceremonies are particularly vulnerable times for virus transmission.

2. Symptoms and sequelae

The early symptoms of Ebola Virus Disease (EVD) are similar to malaria, a disease endemic to West Africa, thereby complicating diagnosis (23). Symptom onset can occur between 2-21 days following exposure (24). Ebola symptoms during the early febrile stage (day 0-3), often includes fever, malaise, body ache, and fatigue. From day 3-10 patients begin experiencing gastrointestinal symptoms including nausea, vomiting, diarrhea, headache, and stomach pain. During days 7-12, a patient either worsens and enters shock or the patient begins to recover. When a patient enters shock the circulatory system fails, leading to organ damage, decreased consciousness, coma, and often death. Alternatively, the patient may begin to improve; diarrhea and vomiting stops, and the patient displays increased energy and improved oral intake. After 10 days patients may develop late complications, which include gastrointestinal hemorrhage (less than 5%), or secondary infection (25). In a report from the largest Ebola Treatment Unit (ETU) in Liberia, approximately 40% of patients began showing symptom improvement around day 10. Nearly all patients who survived until day 13 ultimately lived (25). It may be possible for some individuals to develop mild symptoms and recover (26). Additionally, it is hypothesized that some people may develop asymptomatic infections (26, 27).

3. Previous EVD outbreaks

The largest outbreaks of EVD prior to 2014 have occurred in Central Africa. In 1976 Ebola virus was first identified within Zaire, or modern day Democratic Republic of the Congo (DRC), as well as simultaneously identified in Sudan, or modern day South

Sudan (23). The virus strain in Zaire was later classified as *zaire ebolavirus*; whereas the Sudan strain was later classified as *sudan ebolavirus* (23). It was hypothesized that both of these outbreaks originated from Nzara, South Sudan (26). There are currently five strains of EVD, four of which are deadly to humans (8). EVD was associated with a case fatality of 30-90% depending upon virus strain (8). The Guinean Ebola virus strain showed 97% match in genetic variation to Ebola virus strains from DRC (8).

In the 1976 Sudan outbreak, 284 cases were identified in association with 151 deaths (CFR = 53%) (26). In the Zaire outbreak 318 persons became infected with the disease, resulting in 280 deaths (CFR = 88%) (28). Both epidemics were rapidly stopped once basic isolation techniques, and improvements in handling contaminated materials were implemented (28). The natural reservoir for the virus was not identified (26). Since 1976, DRC has faced six additional EVD outbreaks with the latest occurring in late July 2014. EVD outbreaks within DRC often begin with the handling, butchering, and consumption of raw bush meat (29). Overall, 24 outbreaks of EVD have occurred since it was first identified (23). The largest single outbreak of EVD prior to 2014 occurred in Uganda, during August of 2000. At that time 425 cases including 224 deaths (CFR = 53%) were attributable to Ebola Hemorrhagic Fever (EHF) (30).

4. How the 2014-2015 West African EVD outbreak differs

a. Delay in identification, stigma, and fear

The length of time between initial human exposure to EVD and identification of the disease by authorities (March 2014) allowed for numerous cycles of onwards transmission (8). This delay is partially explainable because EVD had never been identified in West Africa before, and therefore was an unlikely explanatory agent for local healthcare workers (31). However, after the disease was identified, many citizens refused to believe that Ebola was real (32). Rather, many citizens believed that Ebola was caused by witchcraft, and therefore only treatable by a traditional healer (33). Rumors included healthcare workers bringing the disease (34), ETUs harvesting organs for cannibalistic rituals (32), the government lying about EVD to receive international funds (35), and that bleach sprayers contained EVD (36). These rumors not only led people to avoid the health facilities, but also spurred violence towards Ebola responders (35-37). From an ETU study in Liberia, it was estimated that patients did not arrive at the treatment unit until 2-3 days following gastrointestinal symptoms developed at which time infectiousness is high (25). The reasons given for the delay in treatment often include denial of the disease or trust in traditional cures.

b. High mobility and environmental context

The 2014 EVD outbreak began in Meliandou, Guinea located in Guekedou District (Figure 1). Although Meliandou is a remote and sparsely populated village of 31 households, it is located near the district capital of Gueckedou, a regional transportation route. West Africa is characterized by a high degree of population movement, primarily driven by poverty and the search for employment (38). Recent studies estimate that population movement in West Africa to be seven times higher than other places on the globe (38). This high level of mobility within and across borders led to rapid disease transmission within and between nations. Whereas previous EVD outbreaks were geographically confined to a single or few rural regions, many infected individuals during the current outbreak moved between cities and countries. This made contact tracing difficult (38). For reference, on June 21, 2014 Medecins Sans Frontieres (MSF) announced that Ebola patients were identified in over 60 separate locations across the three West African countries (39).

c. Traditional healers

Guinea, Liberia, and Sierra Leone have only recently emerged from years of civil war and conflict that left health infrastructure nearly non-existent (38). Before the 2014 EVD outbreak Liberia had an estimated 1.4 physicians per 100,000 population, Sierra Leone had 2.2 physicians, and Guinea had 100 physicians per 100,000 (40). Limited physician numbers, lack of trust in healthcare workers, as well as limited transportation options to health facilities led people to use alternative health models; primarily traditional healers (34). Traditional healers are well known and respected members of the community. Additionally, they attend to the patient at home, thereby easing transportation needs and changing the expectation of care delivery (34). However, traditional healers were unprepared for EVD care, often undertreating and ignoring quarantine recommendations. Moreover, they were identified as significant transmitters of the disease. Traditional healers would frequently care for EVD patients without protective equipment, often contracting the disease due to successive exposures, and die from Ebola virus infection. Their funerals would result in mass gatherings, where as part of the burial process friends and family would touch the deceased body thereby further propagating the disease. For example, in the early epidemic links one traditional healer funeral with as many as 365 Ebola-related deaths (41).

d. Burial practices

In Guinea 85% of the population is Muslim, and 8% are Christians (42). In Liberia 8% of the population is Muslim and 85% are Christian (43). In Sierra Leone 71%

of the population is Muslim (44). Traditional burial practices differ based upon cultural and religious practice. West African Muslim burial practices include the washing, cleaning, and dressing of the deceased in a favorite outfit. The family and community members then gather for mourning. Women wail and the men dance to demonstrate respect for the deceased; the greater the community respect, the longer the ceremony. As the ceremony is coming to a close, a common bowl is used for ritual hand washing, and a final touch or kiss on the face of the deceased is bestowed. The body is buried on land adjoining the family's home, as the family wants the spirit to not feel forgotten (45). These traditional burial practices do not often result in extended disease transmission for other illnesses. However, Ebola virus remains in the body and body fluids of deceased patients for several days post-mortem, thereby making the washing, touching, and kissing of deceased patients during traditional burial practices a high-risk activity (38). Within Liberia, fear of Ebola led politicians to proscribe traditional burial practices regardless of cause of death, and permit only cremations (46). This policy raised levels of distrust by the local population as not only were their loved ones often taken from the home to ETUs, but now traditional respects to the deceased were denied, and families often had limited idea where the remains of family members were placed (46). These local issues of distrust combined with distrust in Western medicine, led safe and dignified burial teams across West Africa to face problems of escalating violence (35).

5. Case definitions

Case definitions for EVD have been added as affected countries built independent surveillance systems, and as higher case numbers required simpler definitions. The World Health Organization (WHO) case definitions from March to August 2014 are evaluated below. EVD case definitions are divided into three categories: suspected, probable, and confirmed. A suspected case has four criteria; if a patient fulfills any of these criteria, he or she is considered a suspected case. The first includes "any person, alive or dead, suffering or having suffered from a sudden onset of high fever and having had contact with: a suspected, probable or confirmed Ebola case, or a dead or sick animal" (47). Second, a case is suspected if the person presents with a sudden high fever and at least three of the following symptoms: headaches, lethargy, anorexia/loss of appetite, vomiting, diarrhea, stomach pain, aching muscles or joints, difficulty swallowing, breathing difficulties, or hiccups. Third, a suspected Ebola case is any person who presents with inexplicable bleeding. Lastly, a suspected Ebola case includes any sudden or inexplicable death (47).

Probable EVD cases may be identified through two case definitions. A probable case is any suspected case evaluated by a clinician. Second, a probable case is any deceased suspected cases where lab confirmation was not possible and has an epidemiological link with a confirmed case (47). A confirmed EVD case is any suspected case with laboratory confirmation either through positive IgM antibodies directed against ELISA, or detection of virus RNA by reverse transcriptase-polymerase chain reaction (RT- PCR) (47, 48). Research regarding time to accurate negative RT-PCR is mixed; some research supports that RT-PCR should not be relied upon for up to 72 hours after symptoms onset (25, 49), other research states that RT-PCR tests are reliable immediately following symptom onset (50). Any suspected or probable case with a negative laboratory result is considered a non-case (47).

6. Surveillance

On June 25, 2014 the WHO announced that it would only report confirmed cases of EVD at the request of the Sierra Leone Ministry of Health and Sanitation (SLMHS) (51). As a result, Guinea and Liberia epidemiologically report suspected cases, whereas from May 28 to June 24, 2014 WHO reported suspected cases from Sierra Leone and from June 25 to August 28, 2014 the WHO reported confirmed Ebola cases from Sierra Leone. From May 28 to August 29, 2014 SLMHS only reports lab-confirmed cases. As a result, case and mortality reports from May 28 – June 24 do not grow linearly if switching between WHO and SLMHS reports. Additionally, SLMHS was under reporting potential EVD cases, and has different reporting standards than Liberia and Guinea.

7. EVD Conclusion

Numerous cultural, historical, and environmental factors led to the unprecedented case and mortality levels in the 2014-2015 EVD outbreak in West Africa. However, although the outbreak has often exceeded the capacities of epidemiological contact tracing and supportive care, the outbreak has also forced the creation of new technological tools to control the spread of the disease. New advances in contact tracing through mobile data collection either through Short Message Service (SMS) or Interactive Voice Response (IVR) (52, 53), visualization of epidemiological data though interactive and geographically referenced WHO Interactive Map Journals (54), or the development of open data sharing platforms such as the Humanitarian Data Exchange (HDX) (55) are just a few examples of innovative surveillance and reporting systems resulting from this outbreak. The following section will further outline the advances in big data surveillance and how the online news-curating platform, HealthMap, may identify infectious disease outbreaks before traditional surveillance systems.

C. Big data surveillance

1. Introduction

The Internet serves as a living catalog of user behavior, interest, and interactions. Wellman argues that "computer networks are inherently social networks, linking people, organizations, and knowledge" (56, 57). From a public health surveillance perspective, the Internet provides a wealth of information that may be used to better understand health outcomes. Through the network structure, scientists examine the digital records of online interactions including online documents, digital video and photography, purchase metrics, search history, or approval ratings. The following sections will discuss terminology and data collection methods when using online data, as well as the current tools leveraging online media to support public health surveillance.

2. Web 2.0 and health surveillance terminology

Web 2.0 refers to the integration of two-way communication mechanisms on the Internet beginning in the early 2000s. Examples of these discussion platforms included user-developed platforms (ie. wikis and blogs), social networking sites, as well as Really Simple Syndication (RSS) feeds. This marked a fundamental shift in the Internet's purpose from a static platform for pushing information to one built upon integrated user contributions. Clay Shirky, a writer and educator on the social and economic effects of the Internet, discusses the changing media consumption dynamic as a result of Web 2.0.

> *"Media landscape in the 20th century was very good at helping people consume, and we as a result got very good at consuming. Now we have media tools that help us consume, but also produce. We weren't couch potatoes because we liked to be, but rather because it was the only option available. We still like to consume, but also like to create and share"* (58).

This change in interface transformed the Internet from a platform for computer science experts to one built for public interaction, ultimately paving the way for mass contribution.

The term "data mining" refers to the extraction of online data as to build descriptive and predictive models of social interactions (59). Ultimately, data mining is a data collection technique. Within the health field, "e-epidemiology" (60), "infodemiology" (61), "digital epidemiology" (62) or "biosurveillance" (63) refers to its analysis. Mobile and internet-based data collection often has validation measures built into the collection and transcription phases therefore reducing time and cost in study implementation. Nonetheless, new challenges in analyses are introduced from these techniques (62). Currently, three big web-based surveillance types are used to facilitate disease prediction: ProMED-mail, the Public Health Agency of Canada's Global Public Health Intelligence Network (GPHIN), and HealthMap (64).

3. HealthMap: How big data surveillance is used for infectious diseases

HealthMap is an automated information system for monitoring, organizing, and visualizing unstructured reports related to global disease outbreaks by time, location, and infectious agent (Figure 2) (65). It was built upon the hypothesis that informal media may provide an early warning for infectious disease transmission, particularly within under resourced areas (66). In order to test this hypothesis HealthMap harnesses massive amounts of disparate, yet valuable information made available through informal online sources including news sources (e.g. Google News, Moreover, Baidu News, and SOS Info), expertly curated reports (e.g. ProMED Mail, and EuroSurveillance) as well as official reports (e.g. World Health Organization (WHO), GeoSentinel, World Organization for Animal Health (OIE), Food and Agricultural Organization of the United Nations (FAO)) (67, 68). The hypothesis of early disease detection within the news has been repeatedly supported (1-6). Furthermore, decreases in lag-time between first record of an outbreak and public reporting is negatively correlated with freer presses and increased Internet usage. Generally, the expected lag-time in public disease reporting decreases from greater than one month for countries with limited Internet users to one day in countries where over 75% people use the Internet (69).

John Brownstein, developer of HealthMap, has argued that digital epidemiology complements traditional health surveillance (67). Unstructured online reports, such as those encouraged via Web 2.0, are a real-time complement to traditional indicator-based public health disease surveillance methods. Traditional surveillance is limited by wide geographical gaps in coverage and often suffers from difficultly exchanging information across national borders (70). HealthMap helps to fill gaps in locations where traditional surveillance is time, cost, or logistically prohibitive (67). Online sources, by their nature, provide a highly contextual and networked framework at a hyper-local level (ie. publication referenced and time stamped). These characteristics are often necessary to limit the effects of infectious disease spread (57, 71). Yet sometimes digital studies portray a reality different from existing health surveillance models. When this occurs further examination is required in order to determine which methods most validly estimates reality (66).

4. Discussion and future research

There are numerous limitations and biases inherent to big data surveillance. Alessandro Vespignani demonstrated how more data does not necessarily correlate with better outcomes if the information is not appropriate for the questions of interest, thereby expounding the limitation of Google Flu Trends (72). Similarly, Alain-Jacques Valleron of the Pierre and Marie Curie University of Paris states, "the new [infectious disease surveillance] systems depend too much on old existing ones to be able to live without them" (73). Although both good criticism of supplemental disease surveillance HealthMap was not developed for surveillance itself, primarily early warning.

For early warning, HealthMap has demonstrated some potential biases towards reporting from countries with stronger media presence and stronger health systems. In addition, panic induced searches distort the importance of a given health event therefore giving it greater weight within the system (67). Both of these issues are difficult to address within the system itself. Additionally, news curating is cross-sectional in nature, and therefore difficult to reproduce. Although the web tends to provide a time stamped network of information, links may break and search algorithms change with an exponentially growing pool of information (72).

5. Big data surveillance conclusion

The literature is full of research outlining terminology for online surveillance (ie. digital epidemiology, infodemiology, bio surveillance) as well as demonstrating the potential of such systems. However, researchers need to move beyond demonstrating a proof of concept for online surveillance techniques, and continue rigorously testing available models. Most of the current research is quantifying effects, with some researchers moving beyond descriptive statistics to exploratory analyses. However, few researchers have developed inferential or predictive models. Those who have developed predictive models, such as Google Flu Trends, were heavily criticized by the media when the model was no longer accurately forecasting. As data storage is no longer an obstacle for data collection, passive online surveillance will likely continuing growing as a method for improving public health comprehension.

D. Literature Review Conclusion

Research regarding Ebola prior to the 2014 outbreak in West Africa was limited due to financial and probabilistic constraints. However, in the months following the outbreak significantly more information regarding transmission cycles, phylogenetic, treatment and care, as well as health education has been shared with the scientific and public communities. HealthMap first identified Ebola via news reports 9 days prior to official reports (74, 75). As such, the HealthMap system performed in exactly the manner it was designed for. The question of interest now, is above and beyond early warning what else can news articles curated via HealthMap be used for? Bid data surveillance is rife with interrelated terminology, surveillance theory, and small publications demonstrating proof of concept. Only now do researchers have enough financial support to start addressing bigger questions of identifying gaps in surveillance, and fully examining algorithm produced content. This research is interested in the news content produced as “excess” of the HealthMap early infectious disease warning system.

The public is more likely to obtain information regarding Ebola through the news as compared to scientific journals. The manner in which news media portrays scientific information, particularly case and mortality values has an impact upon trust in formal health agencies, public donations, and fear mongering. If the news is consistently over reporting case and mortality numbers, it may cause undue fear in the public and undermine trust in health agencies. If the news is consistently under reporting case and mortality numbers, then it may limit financial support and not convince the public of the situational gravity. More likely news agencies vary in under reporting, over reporting, and accurate reporting thereby giving mixed messages and confusing the public on the true impact of the disease. This study intends to quantify the potential levels of variation between news and official case and mortality reports.

Methods

A. Hypothesis

Previous research has demonstrated that HealthMap can effectively be used as a sentinel surveillance system for early detection of infectious diseases (1-6). However, there is little research exploring potential additional uses of the platform. This study intends to use HealthMap curated news reports as a representative sample of online news to determine the variation between Ebola news reported case and mortality values via HealthMap as compared to the WHO and Ministry of Health (MOH) official reports via the Humanitarian Data Exchange (HDX). The primary question of interest evaluates variation between news and WHO/MOH case and mortality over six months. Furthermore, on average, was the news over or under predicting official reports? The researchers assume delayed count reporting for news. Where official reports are released, then the news duplicates those numbers at a later time. If this were true, then news reports would lag behind WHO/MOH reports and news would either match official reports per day, or news reports reflect older official counts. Alternatively news could receive early reporting and pre-release official numbers. Lastly, news reported values could be scattered randomly around WHO/MOH reports.

B. Study Design

The "Subnational time series data on Ebola cases and deaths in Guinea, Liberia, Sierra Leone, Nigeria, Senegal and Mali since March 2014" was used, and is hereby referenced as "WHO/MOH" collected via the Humanitarian Data Exchange (HDX) (76). This data was manually compiled each day by members of the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) Regional Office of West and Central Africa (ROWCA) from a compilation of WHO, MOH, and other sources (76). This dataset was chosen over the WHO Situational Awareness and the WHO Viral Hemorrhagic Fever (VHF) reports due to open availability of the aggregated data and due to continuous update from the research team. WHO/MOH data was pulled on February 19, 2015 (76).

HealthMap, as referenced previously in the Literature Review, contains news reports related to disease outbreaks (9). The researcher manually collected 13 variables from HealthMap curated news reports. To support data collection an Access database (Microsoft Office Professional Plus 2013) was developed and a form was used for data entry (Figure 3). The data collection form used a nested table design. Logically, each day could contain multiple news reports, and each news reports could contain multiple case or mortality values. As a result, the database structure followed a similar design and was referenced within the form (Figure 3).

Both the HealthMap and HDX data represented serial cross-sectional samples, a common surveillance method. HealthMap was established as a "breaking news" aggregation site for infectious diseases. The tendency for HealthMap algorithms to identify breaking news from a variety of reporting sources is a strength of the platform under its intended purpose, however it raises difficulties in longitudinal study design. This limitation in non-repetitive sources is partially addressed by coding reporting location data, and leveraging this variable as a repeat reference (please see variable selection). The WHO/MOH data, although from significantly fewer reporting sources, also faces challenges in non-repetitive data collection for each day of interest.

C. Variable Selection

1. WHO/MOH

The Excel output was reduced to only include dates from March 23 to August 28, 2014. August 28, 2014 was chosen for conclusion because it was approximately six months after WHO identified EVD in Guinea, and exactly three months after WHO began reporting EVD from within Sierra Leone. Sierra Leone Ministry of Health began reporting on May 25, 2014 and therefore would have a slightly earlier three-month timeline. Sierra Leone was the last country amongst Guinea, Liberia, and Sierra Leone to have EVD identified within its borders. Original WHO/MOH variables included *Country, Category, Sources, Value, Date, Localite,* and *Link* (Appendix 1).

Levels were dropped from *Country, Category,* and *Sources* and the results were numerically re-coded. For the variable *Country,* "Nigeria" was excluded as it was located outside of the study area of interest. "Confirmed Cases," "Probable Cases," and "Suspected Cases" were dropped from *Category*. The frequency of "Cases" during the study duration (March 23 - August 28, 2014) within the three countries of interest (excluding Nigeria) approximated the summed frequency of "Confirmed Cases," "Probable Cases," and "Suspected Cases" when compared across the same time period and locations. The researcher chose to keep "Cases" as representative. *Sources* were collapsed into "International" agencies such as WHO and UNICEF and "National" agencies such as MOH. *Localite* contained the district name for 64 districts across the three countries of interest as well as a "National" total. "National" was chosen as representative for *Localite*. If a given *Date* did not have a "National" code, one was created. Sierra Leone was the only *Country* with differences between the district sum and “National” totals. Once national totals were identified *Localite* was dropped from analysis. For more information regarding HDX variable use, please see Appendix 1: Data Dictionary: HDX Sub-national time series (March 23 – August 28, 2015)

2. HealthMap

Thirteen variables from HealthMap news reports were collected using a Microsoft Access form, excluding three unique identifiers representing the three data tables (*HMID, ReportID, and ValueID*) (Appendix 2). The remaining variables included *Date*, news agency name (*News Agency)*, news agency reporting location (*Location*), *Notes*, reason for missing counts (*Reason Missing*), *URL*, *Country*, type of counts (*Category*), *Value* (also known as count), agency providing counts (*Reference*), whether a person was named in reference of the counts (*Named*), whether only Guinea, Liberia, and Sierra Leone were estimated in West African totals (*Not Adj*), or whether the count was approximated (*Approx*). Figure 3 shows the Microsoft Access data collection form for March 14, 2014.

Initial data collection occurred from December 15 - 27, 2014. Secondary data collection and revision of qualitative codes was completed from January 19 - February 1, 2015. During secondary data review some HealthMap articles previously viewed in primary data collection were no longer available. In these cases a cached version of the website were visited when possible. Once all data was collected, qualitatively coded, and cleaned, then missing HealthMap reports due to broken URL links were searched using Google. If the article title, date, and article abstract referenced on HealthMap matched the Google search results, then the values were collected. Articles identified through Google search were specially designated as such within the database. Only national estimates were collected for new cases, cases, deaths, and new deaths. Breaking news articles regarding the death of important figures were not included in "new deaths" totals. Articles only containing individual or local cases/deaths were identified as missing for the purposes of this study. News articles referencing approximated values, for example "approximately 500 people have died in Guinea" were designed as such (*Approx.*).

3. WHO/MOH and HealthMap merged dataset

Dataset, category, and value variables were combined to create new variables for analysis (ie. *HM_Cases*, *HDX_Deaths*). As either many or no news reports existed each day, WHO/MOH case and mortality reports were matched to HealthMap reports by country and date. All non-matched WHO/MOH reports were removed from analysis. As normality and constant variance assumptions were reviewed, additional calculations including log-transformation and differences between WHO/MOH and news-curated case and mortality estimates were constructed. As neither of these statistical transformations resulted in non-statistically significant assumption test, the non-transformed data was used. Additional calculations regarding log-transformations for HealthMap and WHO/MOH case and mortality values, as well as calculated differences were included within the final dataset. Final variables for inclusion in the combined dataset are outlined in Appendix 3: WHO/MOH and HealthMap merge

D. Method of Analysis

Regression was used to evaluate the relationship between WHO/MOH case and mortality estimates separately with HealthMap news derived variables. Collinearity was assessed prior to model selection. A Variance Inflation Factor (VIF) greater than 5 or Condition Index (CI) greater than 30 were cutoffs for variable removal during the collinearity assessment. Calendar date resulted in a collinearity problem, so *Date* was reoriented to grow sequentially based upon date of first case identification (e.g. Guinea, March 14, 2014 = 1, Guinea March 19, 2014 = 6). Furthermore, the variable *Approximate* was removed from assessment for cases due to collinearity (n=3). No further collinearity issues were identified.

Two sets of models were derived from the data. Both models reviewed the following HealthMap variables: HealthMap values, country, reoriented date, news agency location, news agency size, approximated count, and reference (sub-analysis). The "full" data analysis reviewed all variables excluding *References* for cases (Model 1) and deaths (Model 2). A "sub-variable analysis", which included *References,* was further reviewed using the same variables for cases (Model 3) and deaths (Model 4). The sub-analyses were developed because of the high number of missing values (n=99) associated with referenced data sources (Appendix 4). For review of the model selection tables please see Appendix 3 and 4. Models compared WHO/MOH counts to news derived or news-matched variables. Model selection was determined by comparing forward, backward, and stepwise selection models as well as the full model estimates primarily evaluating Mallow's C(p).

Results

From March 14 to August 28, 2014, 739 case and mortality counts were collected from 279 HealthMap curated news reports. During that time period HealthMap identified news reports for 119 days with 48 days having no associated news report. On average, 2.34 Ebola related news reports were identified per HealthMap reporting day or 1.67 news reports over the entire duration. The range in potentially identified case and mortality counts per news report covered the minimum to maximum (range = 0-16), identified from the four regions (Guinea, Liberia, Sierra Leone, and West Africa) interacting with the four case statuses (Cases, Deaths, New cases, and New deaths) (Table 1). News reports on July 8 and July 18 contained the maximum number of HealthMap case and mortality counts. On average, 2.64 case and/or mortality estimates were provided for each news report. Through data cleaning 77 news reports, representing 21 unique reporting days, were removed due to not containing counts (Table 2). Additionally, 339 case and mortality counts were removed due to representing regional totals (n= 207), incidence estimates (n=113), or unverified rumors of Ebola in country (n=20) (Table 1). This resulted in 400 remaining case (n=178) and mortality (n=222) counts to be evaluated. These estimates are derived from 202 HealthMap curated news reports over 98 days, or 72% of all HealthMap curated news reports from March 14 to August 28, 2014.

As noted in the literature (74, 75), news reports identified Ebola within each of the three countries of interest prior to official reports (Table 3). Guinea was first identified on March 14, 2014 with the first official report 9 days later on March 23, 2014. Cases were first identified in the news for Liberia on March 24, 2014 with the first official report released on April 10, 2014; 17 days after it was first identified in the news. For Sierra Leone, rumors of the disease began circulating in the news on March 22, 2014. On March 25, 2014 the news began reporting cases, and on March 28, 2014 the first official report was released, one week after it was first identified (Table 3).

Figures 4-7 visually depict the case and mortality counts via the news and WHO/MOH estimates for Guinea (Figure 4), Liberia (Figure 5), Sierra Leone (Figure 6) and all combined countries (Figure 7). News reported case and mortality peaks in the first month following the initial news report and again for one month starting mid-June (Figure 4-7). For Guinea, the most new reported counts occur one month following March 14 (n= 74), and again for the month of June (n= 44) (Figure 4). Liberia shows a similar pattern, with a high number of news reports for one month following March 24 (n= 40), and again for the month of July (n= 68) (Figure 5). The news reporting peaks for Guinea and Liberia in June, are associated with the peak in Sierra Leone news reporting starting in June (n= 68) and continuing through July (n= 43) (Figure 6). Figure 7 overlays are three countries WHO/MOH and news reported case and mortality counts from March 14 – August 28, 2014, and demonstrates the surge in news reported counts in June and July (n=272). These two months account for 68% of the collected news-derived counts.

Table 4 provides variable counts from HealthMap news articles. Within the news, Guinea had the most identified case (n=65) and mortality (n=81) counts, followed by Sierra Leone (cases=58, deaths=71) and Liberia (cases=55, deaths=70). Mortality estimates (n=222) were more prevalent than case estimates (n=178) across all three countries, which reflect a shift to only report mortality estimates as the number of infected increases. International agencies such as the WHO were cited 171 times (cases=85, deaths=86) as the definitive source for reported counts, whereas local agencies such as the MOH were referenced 130 times (cases=60, deaths=70). Article references had a high number of missing values (missing=99), resulting in the need for a sub-analysis to further evaluate the relation. Additional covariates included approximated values (cases=3, deaths=19), reporting from an affected country (cases=66, deaths=89), and reporting via major news agency (cases=59, deaths=92).

We matched news reported case and mortality counts by day and country with official WHO/MOH counts from March 14 to August 28, 2014 and compared the mean counts (Table 5). The mean news-derived count for Ebola cases in Guinea was 233. This was higher than the mean count for cases in Sierra Leone (mean=209) and Liberia (mean=158). Similarly, the mean count for deaths was higher in Guinea (mean=242) then in Sierra Leone (mean=210) and Liberia (mean=209). This ordered ranking for higher case and mortality estimates (Guinea, Sierra Leone, then Liberia) mimics the decreasing number of news reports over the same time period (Guinea, Sierra Leone, then Liberia) (Table 4). In comparing mean counts between news and WHO/MOH, the agency reporting higher mean counts was not consistent across countries. In Guinea the WHO/MOH reported higher mean counts. In Liberia, the news reported higher mean counts. Whereas in Sierra Leone there was little difference between WHO/MOH and news counts. The greatest daily variability in case counts was within Liberia (mean difference=55, sum of differences=2,034). For deaths, the mean daily difference was greatest in Sierra Leone (mean difference=26, sum of differences=1,830).

In reviewing the reported data sources' case and mortality counts, official WHO/MOH estimates were more likely to cite an international agency such as the WHO (cases=120, deaths=147) as compared to a local agency such as the MOH (cases=34, deaths=42). News reports also were more likely to cite sources internationally (cases=85, deaths=86), then locally (cases=60, deaths=70). Although more data was referenced internationally, the numbers of citations are closer within the News (cases difference=15, deaths difference=16), then within the official estimates (cases difference=86, deaths difference=105).

Table 6 shows the estimates from the full regression models for cases and deaths (Model 1 and 2). Model 1 and 2 both resulted in the same variable selection: news derived counts, date (reoriented), and country (Table 6). Model assessment for the cases sub-analysis (Model 3) was not unanimous. Forward and Stepwise selection identified two variables for inclusion (news derived case values and reporting location). Backwards elimination identified four variables for inclusion (news derived case values, reporting location, date, and country). The full model analysis supported the same four variable model identified in backwards elimination (Mallow's C(p) = 4.64), although a five variable model including news agency size was also supported (Mallow's C(p) = 5.30). Ultimately the four variable model was selected based upon Mallow's C(p). All model selection techniques for the mortality sub-analysis supported a three variable model (Model 4). Selected variables for inclusion were: news derived death values, country, and referenced data source. In evaluating the beta estimates, there was not an exact one to one relationship between news and WHO/MOH derived counts, which would be expected if news was an identical copy of official estimates (cases t=-28.17, p<0.0001; deaths t= -10.11, p<0.0001).

Table 6 shows the t-test, ANOVA, and beta estimates for the case and mortality outcome full analyses (Model 1 and 2) these models include all available data. The full analysis case model predicted 75.9% of the change in estimate for WHO/MOH cases and 90.3% for deaths. This implies that 24.1% of the variation for cases, and 9.7% of the variation in mortality estimates cannot be explained by variables identified within the news. For every increase in news derived cases, WHO/MOH cases increase by 2.79 when controlling for reoriented date and county (news case beta=0.36, p<0.0001). For every new news derived death, WHO/MOH deaths increase by 1.45 when controlling for reoriented date and country (news deaths beta=0.69, p<0.0001). Both estimates were statistically significant. Each day, the number of WHO/MOH cases increases by 0.46 (date cases beta=2.17, p<0.0001) and WHO/MOH deaths increase by 1.24 (date deaths beta=0.81, p<0.0001) when controlling for news derived values and country. As a result, WHO/MOH cases estimates are growing faster than mortality estimates, however on a per day basis WHO/MOH mortality was increasing faster than case identification.

Table 7 shows the t-test, ANOVA, and beta estimates for the case and mortality outcome sub-analyses (Model 3 and 4) these models include the referenced data source variable, which had high missing. In evaluating the beta estimates for the model of cases, we fail to reject the null hypothesis that the slope for news equals 1 and therefore conclude that WHO/MOH derived and news derived counts are not statistically different (t=-1.18, p=0.2399). For deaths, we reject the null hypothesis that the slope for news equals 1 and therefore conclude that WHO/MOH and news derived counts are statistically different (t=7.71, p<0.0001). The sub-analyses model predicts 95.9% of the variation in WHO/MOH deaths, implying that 4.1% of the variation in mortality cannot be explained via the model. For every new WHO/MOH death, news derived mortality increases by 1.17 when controlling for county and referenced data source (news deaths beta=1.17, p<0.0001). When the MOH is referenced, the death toll on average will be 13.75 counts higher than WHO estimates.

Discussion

As demonstrated previously in the literature, news reports first identified Ebola within the countries of interest prior to official reports (Table 3). As with other machine learning algorithms, public hype regarding an event will change the weighed importance of the event within the system. Identifying ways to maintain and strengthen classification as to further discern signal from noise in online surveillance will be a continuing challenge. In evaluating the news-derived prediction models, there was not an exact relationship between news and WHO/MOH derived counts as to be expected if news was an identical copy of official estimates. For both cases and deaths, news-derived counts, reoriented date, and country were significant for model inclusion. The full analysis case model predicted 75.9% of the change in estimate for WHO/MOH cases and 90.3% for deaths. For mortality estimates, given that three variables (news counts, reoriented date, and country) can accurately predict over 90% of the variation in reporting, supports general accuracy in media mortality reporting. Nonetheless, this also implies that 24.1% of the variation for cases, and 9.7% of the variation in mortality estimates cannot be explained by basic variables identified within the news. It is hypothesized that the lower predictive capacity for cases was related to the greater variability in case estimates due to multiple case definitions (suspected, probable, and confirmed cases) as well as changes in case reporting for Sierra Leone during the time of study. Both of these factors could make reporting more difficult for news agencies and therefore result in misattribution of case numbers.

When narrowing the inclusion criteria to news reports containing references for case and mortality counts, there was no statistically significant difference in reporting between news and official reports for cases. However, there was a statistically significant difference for deaths. The sub-analysis model, which included news-derived mortality, estimates, reoriented date, country, and references, predicted 95.9% of the change in estimates for deaths. This resulted in 4.1% of the variation in mortality reports unexplained via news. The difference in new-derived variable significance for cases and deaths, as well as the high prediction model for mortality implies two things. First, cases and deaths should be analyzed together as to quantify the impact of reporting type. Second, that when a news article cites it source for case and mortality counts, then on average the value will be accurate.

A. Strengths and Weaknesses

Two strengths of this study included evaluating the potential secondary use of HealthMap within online surveillance, as well as quantifying the potential for distrust in media reported values where there was limited research prior. However, this study was limited in design by non-repetitive news report sources and original longitudinal design, which led to difficultly identifying an appropriate statistical approach in the original form. Additionally, the Humanitarian Data Exchange (HDX) subnational time series provided the best publicly available estimate of case and mortality during that time period, however it is not the most accurate dataset available. As a result, differences in case and mortality report may not be due to the news source, but rather be as a result of inaccuracies in the official report.

Future Directions

Numerous alternative analyses could be completed using the data collected from HealthMap. Specifically, five alternative analyses are highlighted: case and death comparisons, rumors, regional-level reporting, longitudinal time series as well as incidence analysis. As mentioned in the Discussion section, cases and deaths could be analyzed within the same model. It was assumed that mortality estimates were a subset of case estimates and therefore acceptable to review independently. However, the sub-analysis suggested otherwise. The data should be recoded to evaluate reporting type (case or deaths) as another variable of interest. Second, 20 rumors of Ebola were identified via the news from March 14 – August 28, 2014. This included 17 rumors regarding Ebola cases and 3 rumors regarding Ebola deaths ranging in date from March 14 to August 11. Further analyses should evaluate why rumors of Ebola perpetuated in the news long after the disease was identified in each country. Third, researchers should compare country-level reporting to West Africa regional reporting. Although not highlighted within this analysis, country-specific reporting was limited after July and regional reporting became prevalent after mid-June (Figure 8). Future research should determine why this shift from country to regional reporting occurred, and the potential dissociative impact. Fourth, although this research evaluated the impact of news using a continuous date, additional research could use a longitudinal design as to better quantify how case and mortality reporting changes over time. Lastly, new cases (n=59) and new deaths (n=223) were collected via HealthMap news reports. A similar analysis could evaluate how new cases and new deaths compare to WHO/MOH incidence estimates during that same time period.

Beyond spurring additional research questions, this research evaluated potential secondary uses of HealthMap curated news reports. Specifically, the new-curated reports were used as a representative sample of all online news to compare variation in unofficial and official reporting for Ebola case and mortality. Although similar variability analyses could be completed with other infectious diseases, the disease would need to be "news worthy" enough to spur articles over time. MERS, US measles, or a specific influenza strain may have enough articles to develop similar analyses. Overall, news curated reports work well for identifying first emergence of a disease, however after a given duration news is often time delayed official reporting. Research such as this may not help health responders directly, but rather evaluates public perception of a disease, which in turn could be used to evaluate potential financial or political support. Such a model should not be used for predicting official estimates. Rather news articles should directly link news consumers with official reports; therefore reducing variability in the public domain.

References

1. Brownstein JS. Healthmap: Global Disease AlertMapping. 2006.
2. Hay SI, George DB, Moyes CL, et al. Big data opportunities for global infectious disease surveillance. *PLOS Medicine* 2013;10(4):4.
3. Bennett KJ, Olsen JM, Harris S, et al. The Perfect Storm of Information: Combining Traditional and Non-Traditional Data Sources for Public Health Situationa Awareness During Hurricane Response. *PLOS Current Disasters* 2013.
4. Chan EH, Sahai V, Conrad C, et al. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS neglected tropical diseases* 2011;5(5):e1206.
5. Brownstein JS, Freifeld CC, Madoff LC. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. *The New England Journal of Medicine* 2009;360(21):5.
6. Lau EH, Zheng J, Tsang TK, et al. Accuracy of epidemiological inferences based on publicly available information: retrospective comparative analysis of line lists of human cases infected with influenza A(H7N9) in China. *BMC medicine* 2014;12:88.
7. World Health Organization Ebola Response Team. Ebola virus disease in West Africa--the first 9 months of the epidemic and forward projections. *N Engl J Med* 2014;371(16):1481-95.
8. Baize S, Pannetier D, Oestereich L, et al. Emergence of Zaire Ebola virus disease in Guinea. *N Engl J Med* 2014;371(15):1418-25.
9. HealthMap. 2014 Ebola Outbreaks. Online; 2014. (http://www.healthmap.org/ebola). (Accessed November 1 2014).
10. World Health Organization. Ebola virus disease In Guinea. Online: WHO; 2014. (http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4063-ebola-hemorrhagic-fever-in-guinea.html). (Accessed February 7 2015).
11. World Health Organization. Statement on the 1st meeting of the IHR Emergency Committee on the 2014 Ebola outbreak in West Africa. 2014. (http://www.who.int/mediacentre/news/statements/2014/ebola-20140808/en/). (Accessed February 8 2015).
12. World Health Organization. Ebola virus disease, West Africa - update (28 March 2014). 2014. (http://www.who.int/csr/don/2014_05_28_ebola/en/). (Accessed February 15 2015).
13. World Health Organization. Ebola Virus Disease, West Africa (28 August 2014). Online: WHO; 2014. (http://www.who.int/csr/don/2014_08_28_ebola/en/). (Accessed February 15 2015).
14. Mari Saez A, Weiss S, Nowak K, et al. Investigating the zoonotic origin of the West African Ebola epidemic. *EMBO molecular medicine* 2014;7(1):17-23.

15. Centers for Disease Control and Prevention (CDC). Facts about Bushmeat and Ebola. 2014. (http://www.cdc.gov/vhf/ebola/pdf/bushmeat-and-ebola.pdf). (Accessed February 8 2015).
16. Alssatou B. Santé : Une étrange fièvre se déclare à Macenta, plusieurs cas de morts signalés.... Africaguinee.com; 2014. (http://www.africaguinee.com/articles/2014/03/14/sante-une-etrange-fievre-se-declare-macenta-plusieurs-cas-de-morts-signales). (Accessed December 15 2014).
17. Centers for Disease Control and Prevention (CDC). Review of Human-to-Human Transmission of Ebola Virus. Online: CDC; 2014. (http://www.cdc.gov/vhf/ebola/transmission/human-transmission.html). (Accessed February 5 2015).
18. Rodriguez LL, De Roo A Fau - Guimard Y, Guimard Y Fau - Trappier SG, et al. Persistence and genetic stability of Ebola virus during the outbreak in Kikwit, Democratic Republic of the Congo, 1995. (0022-1899 (Print)).
19. World Health Organization. Interim advice on the sexual transmission of the Ebola virus disease. 2015. (http://www.who.int/reproductivehealth/topics/rtis/ebola-virus-semen/en/). (Accessed July 3 2015).
20. Varkey JB, Shantha JG, Crozier I, et al. Persistence of Ebola Virus in Ocular Fluid during Convalescence. *New England Journal of Medicine* 2015;372(25):2423-7.
21. Bausch DG, Towner JS, Dowell SF, et al. Assessment of the Risk of Ebola Virus Transmission from Bodily Fluids and Fomites. *The Journal of Infectious Diseases* 2007;196(Supplement 2): S142-S7.
22. Rowe AK, Bertolli J Fau - Khan AS, Khan As Fau - Mukunu R, et al. Clinical, virologic, and immunologic follow-up of convalescent Ebola hemorrhagic fever patients and their household contacts, Kikwit, Democratic Republic of the Congo. Commission de Lutte contre les Epidemies a Kikwit. (0022-1899 (Print)).
23. World Health Organization. Ebola Virus Disease (Fact Sheet N. 103). 2014. (http://www.who.int/mediacentre/factsheets/fs103/en/ -). (Accessed October 9 2014).
24. World Health Organization. Frequently asked questions on Ebola virus disease. 2014. (http://www.who.int/csr/disease/ebola/ebola-faq-en.pdf?ua=1). (Accessed October 9, 2014 2014).
25. Chertow DS, Kleine C Fau - Edwards JK, Edwards Jk Fau - Scaini R, et al. Ebola virus disease in West Africa--clinical manifestations and management. 2014(1533-4406 (Electronic)).
26. Report of a World Health Organization International Study Team. Ebola haemorrhagic fever in Sudan, 1976. *Bulletin of the World Health Organization* 1978;56(2):247-70.
27. Leroy EM, Baize S Fau - Volchkov VE, Volchkov Ve Fau - Fisher-Hoch SP, et al. Human asymptomatic Ebola infection and strong inflammatory response. 2014(0140-6736 (Print)).

28. Heymann Dl Fau - Weisfeld JS, Weisfeld Js Fau - Webb PA, Webb Pa Fau - Johnson KM, et al. Ebola hemorrhagic fever: Tandala, Zaire, 1977-1978. (0022-1899 (Print)).
29. World Health Organization. Democratic Republic of Congo: “classic” Ebola in a country experiencing its seventh outbreak. 2014. (http://www.who.int/csr/disease/ebola/ebola-6-months/drc/en/). (Accessed February 8 2015).
30. Centers for Disease Control and Prevention (CDC). Outbreak of Ebola Hemorrhagic Fever ---Uganda, August 2000--January 2001. *MMWR* 2001;50(5):73-7.
31. World Health Organization. Ground zero in Guinea: the outbreak smoulders - undected - for more than 3 months. 2014. (http://www.who.int/csr/disease/ebola/ebola-6-months/guinea/en/ -). (Accessed January 21 2015).
32. Estrada C. Ebola, snakes and witchcraft: Stopping the deadly disease in its tracks in West Africa - International Federation of Red Cross and Red Crescent Societies; 2014. (http://www.ifrc.org/en/news-and-media/news-stories/africa/sierra-leone/ebola-snakes-and-witchcraft-stopping-the-deadly-disease-in-its-tracks-in-west-africa-66215/). (Accessed February 9 2015).
33. Freeman C. Ebola outbreak: fight against disease hampered by belief in witchcraft, warns British doctor. 2014. (http://www.telegraph.co.uk/news/worldnews/africaandindianocean/sierraleone/11001610/Ebola-outbreak-fight-against-disease-hampered-by-belief-in-witchcraft-warns-British-doctor.html). (Accessed February 8 2015).
34. Ferme M. Hospital Diaries: Experiences with Public Health in Sierra Leone. 2014. (http://www.culanth.org/fieldsights/591-hospital-diaries-experiences-with-public-health-in-sierra-leone). (Accessed February 8 2015).
35. Doe C. Ebola Outbreak Feeds on Fear, Anger, Rumors. ABC News; 2014. (http://abcnews.go.com/Health/ebola-outbreak-feeds-fear-anger-rumors/story?id=24822436). (Accessed February 9 2015).
36. Reuters. Guinean Security Forces Break Up Riot in Ebola-racked South. Voice of America; 2014. (http://www.voanews.com/content/reu-guinean-security-forces-break-up-riots-in-ebola-racked-south/2431976.html). (Accessed February 9 2015).
37. Gbandia S. Sierra Leone Police Use Tear Gas to Curb Ebola-Related Riot. Bloomberg Business; 2014. (http://www.bloomberg.com/news/articles/2014-07-25/sierra-leone-police-use-tear-gas-to-curb-ebola-related-riot). (Accessed February 9 2015).
38. World Health Organization. One year into the Ebola epidemic: a deadly, tenacious and unforgiving virus. 2015, (Global Alert and Response (GAR)
39. Medecins Sans Frontieres. Ebola in West Africa: Epidemic requires massive deployment of resources. 2014. (http://www.msf.org/article/ebola-west-africa-epidemic-requires-massive-deployment-resources). (Accessed February 9 2015).

40. Christensen J. 'Out of control': How the world reacted as Ebola spread. CNN; 2014. (http://www.cnn.com/interactive/2014/11/health/ebola-outbreak-timeline/). (Accessed February 9 2015).

41. World Health Organization. Sierra Leone: a traditional healer and a funeral. 2014. (http://www.who.int/csr/disease/ebola/ebola-6-months/sierra-leone/en/). (Accessed February 9 2015).

42. US State Department. Guinea 2012 International Religious Freedon Report 2012. (http://www.state.gov/documents/organization/208368.pdf). (Accessed March 28 2015).

43. Bureau of Democracy HR, and Labor. International Religious Freedom Report 2010. 2010. (http://www.state.gov/j/drl/rls/irf/2010/148698.htm). (Accessed March 28 2015).

44. Pew Research Center. Mapping the Global Muslim Population: A Report on the Size and Distribution of the World's Muslim Population. 2009. (http://www.pewforum.org/files/2009/10/Muslimpopulation.pdf). (Accessed March 28 2015).

45. Haglage A. Kissing the Corpses in Ebola Country. The Daily Beast; 2014. (http://www.thedailybeast.com/articles/2014/08/13/kissing-the-corpses-in-ebola-country.html). (Accessed February 12 2015).

46. Baker A. Liberia Burns its Bodies as Ebola Fears Run Rampant. Time; 2014. (http://time.com/3478238/ebola-liberia-burials-cremation-burned/). (Accessed February 21 2015).

47. World Health Organization. Case definition recommendations for Ebola or Marburg Virus Diseases (09 August 2014). 2014,

48. Ksiazek TG, West CP, Rollin PE, et al. ELISA for the Detection of Antibodies to Ebola Viruses. *Journal of Infectious Diseases* 1999;179(Supplement 1):S192-S8.

49. Towner JS, Rollin Pe Fau - Bausch DG, Bausch Dg Fau - Sanchez A, et al. Rapid diagnosis of Ebola hemorrhagic fever by reverse transcription-PCR in an outbreak setting and assessment of patient viral load as a predictor of outcome. (0022-538X (Print)).

50. Ksiazek TG, Rollin Pe Fau - Williams AJ, Williams Aj Fau - Bressler DS, et al. Clinical virology of Ebola hemorrhagic fever (EHF): virus, virus antigen, and IgG and IgM antibody findings among EHF patients in Kikwit, Democratic Republic of the Congo, 1995. (0022-1899 (Print)).

51. Schemm P. New WHO reporting method for Ebola reduces death toll in Sierra Leone. Online: CTV News; 2014. (http://www.ctvnews.ca/health/new-who-reporting-method-for-ebola-reduces-death-toll-in-sierra-leone-1.1885628). (Accessed February 8 2015).

52. UNFPA. Guinea-Conakry: The Introduction of the Mobile Application "CommCare" in Ebola Contact Tracing and Information Management Related to Patients. 2014. (http://wcaro.unfpa.org/public/lang/en/pid/18732;jsessionid=B012D19CEE64B6B77A1AB98B0C3150FB.jahia01). (Accessed February 21 2015).

53. Hermann C. What is ACT? : Assisted Contact Tracing (ACT); 2014. (https://docs.google.com/document/d/1uOxuUr-

LHgAjmEB1FP1kmyoQadCFjXw9N9L4yM7tCBE/edit?usp=sharing). (Accessed February 21 2015).

54. World Health Organization. Interactive Map Journal. 2015. (http://maps.who.int/MapJournal/?appid=7462923f64ef44cb8a1dbde6cda64906). (Accessed February 21 2015).

55. Humanitarian Data Exchange (HDX). West Africa: Ebola Outbreak. 2014. (https://data.hdx.rwlabs.org/ebola). (Accessed November 19 2014).

56. Wellman B. Computer Networks as Social Networks. *Science* 2001;293(5537):4.

57. Anderson JG. Evaluation in health informatics: social network analysis. *Computeres in Biology and Medicion* 2002;32:15.

58. Shirky C. How congitive surplus will change the world. *Ted@Cannes*. Cannes, France, 2010.

59. Srivastava J, Admad MA, Pathak N, et al. Data Mining Based Social Network Analysis from Online Behavior. Presented at Society for Industrial and Applied Mathematics, San Diego, CA2008.

60. Ekman A, Litton JE. New times, new needs; e-epidemiology. *Eur J Epidemiol* 2007;22(5):285-92.

61. Eysenbach G. Infodemiology and Infoveillance: Framework for an Emergeing Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research* 2009;11(1):14.

62. Salathe M, Bengtsson L, Bodnar TJ, et al. Digital Epidemiology. *PLOS Computational Biology* 2012;8(7):5.

63. van Gelder M, Pijpe A. E-epidemiology: a comprehensive update. *OA Epidemiology* 2013;4(1):5.

64. Barclay E. Predicting the next pandemic. *The Lancet* 2008;372(9643):1025-6.

65. Freifeld CC, Mandl KD, Reis BY, et al. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc* 2008;15(2):150-7.

66. Brownstein J, Freifeld CC, Reis BY, et al. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. 2008;5(7):6.

67. Brownstein JS, Mekaru S, Tomasula AF. Digital Epidemiology and Infectious Disease Outbreaks. *Harvard Health Policy Review* 2013;15(1):2.

68. HealthMap. About. HealthMap. (Accessed July 5 2015).

69. McAlarnen L, Smith K, Brownstein JS, et al. Internet and free press are associated with reduced lags in global outbreak reporting. *PLoS Curr* 2014;6.

70. Mayor S. Internet crawler uses unconventional information sources to track infectious disease outbreaks. *Bmj* 2008;337.

71. Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Comput Biol* 2011;7(10):e1002199.

72. David Lazer, Ryan Kennedy, Gary King, et al. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 2014;343:1203-5.

73. Butler D. When google got flu wrong. *Nature* 2013;494(155):2.

74. Online tool nailed Ebola epidemic. Associated Press; 2014. (http://www.politico.com/story/2014/08/healthmap-ebola-outbreak-109881.html?hp=l8). (Accessed July 12 2015).

75. Gilpin L. How an algorithm detected the Ebola outbreak a week early, and what it could do next. TechRepublic; 2014. (http://www.techrepublic.com/article/how-an-algorithm-detected-the-ebola-outbreak-a-week-early-and-what-it-could-do-next/). (Accessed July 12 2015).

76. OCHA ROWCA. Sub-national time series data on Ebola cases and deaths in Guinea, Liberia, Sierra Leone, Nigeria, Senegal and Mali since March 2014. In: HDX, ed, 2014.

77. Figure 2: Map of Location 0, Meliandou Guinea. 2015.

78. HealthMap. Alerts from past week. 2015. (http://www.healthmap.org/en/). (Accessed July 5 2015).

79. Moore R. Dataset: Cleaned HDX Subnational time series data for Ebola news & WHO/MOH variability analysis. http://dxdoiorg/106084/m9figshare1496611. Figshare, 2015.

80. Moore R. Dataset: Cleaned HealthMap Ebola curated news data for Ebola news & WHO/MOH variability analysis. http://dxdoiorg/106084/m9figshare1496612. Figshare, 2015.

81. Moore R. Dataset: Merged cleaned HDX Subnational time series and HealthMap Ebola data for Ebola news & WHO/MOH variability analysis. http://dxdoiorg/106084/m9figshare1496613. Figshare, 2015.

Tables

Table 1: Count of news-curated case and mortality by country (March 14 - August 28, 2014)

	Guinea	Liberia	Sierra Leone	West Africa	Total
Cases	67	60	68	63	**258**
Deaths	81	71	73	111	**336**
New Cases	24	20	15	15	**74**
New Deaths	20	20	13	18	**71**
Total	**192**	**170**	**168**	**207**	**739**

Table 2: News reports excluded by topic due to missing case and mortality counts (March 14 - August 28, 2014)

	Reports Excluded (%)
WHO report	17 (22.08%)
URL	12 (15.58%)
Movement restriction or security incident	10 (12.99%)
Fear or rumor	6 (41.56%)
Unclassified	32 (41.56%)
Total	**77 (100%)**

Table 3: Date of first Ebola news rumor, news report, and official report for Guinea, Liberia, and Sierra Leone (March 14-August 28, 2014)

	Guinea	Liberia	Sierra Leone
Cases			
Ebola news rumor**	--	3/31/2014	3/22/2014*
Ebola news report	3/19/2014	3/25/2014	3/25/2014
Ebola official report	3/24/2014	4/10/2014	3/28/2014
Lag between news (rumor or report) & official reports	6 days	17 days	7 days
Deaths			
Ebola news rumor**	--	7/31/2014	3/24/2014
Ebola news report	3/14/2014*	3/24/2014*	3/26/2014
Ebola official report	3/24/2014	4/10/2014	3/28/2014
Lag time between news & official reports	11 days	18 days	5 days

* First report by country

** Rumored cases (n=17), rumored deaths (n=3)

Table 4: News-curated variable counts by cases and deaths (March 14 - August 28, 2014)

	Cases (n=178)	Deaths (n=222)	Total (n=400)
Country			
Guinea	65	81	146
Liberia	55	70	125
Sierra Leone	58	71	129
*Referenced data sources**			
International agency	85	86	171
Local agency	60	70	130
Approximated Values			
Yes	3	19	22
No	175	203	378
Reporting Location			
Affected Country	66	89	155
Non-affected County	112	133	245
News Agency Size			
Major News Agency	59	92	151
Minor News Agency	119	130	249

* Missing Cases (29), Missing Deaths (66)

Table 5: Unadjusted means for news-curated variables by cases and deaths (March 14 - August 28, 2014)

	Cases (n=178)			Deaths (n=222)		
	News	WHO/ MOH	Difference (sum)	News	WHO/ MOH	Difference (sum)
Country						
Guinea	233	242	22 (1,401)	157	162	17 (1,370)
Liberia	158	134	55 (3,034)	93	97	20 (1,373)
Sierra Leone	209	210	43 (2,486)	78	86	26 (1,830)
Referenced data sources						
International agency	85	120	35	86	147	61
Local agency	60	34	26	70	42	28
Missing	33	24	9	66	33	33

Table 6: Full analysis t-test, ANOVA, and beta estimates for WHO/MOH cases (n=178) and deaths (n=222)

	Cases WHO/MOH (n=178)		Deaths WHO/MOH (n=222)	
	Estimate (CI)	p-value	Estimate (CI)	p-value
T-value*	-28.17	<0.0001	-10.11	<0.0001
F-value	136.53	<0.0001	504.80	<0.0001
R Square	75.9%		90.3%	
Beta Estimates				
Intercept	-56.81 (-76.54- -37.07)	0.0045	-41.96 (-50.30- -33.62)	<0.0001
News	0.36 (0.32-0.39)	<0.0001	0.69 (0.66-0.72)	<0.0001
Date	2.17 (1.96-2.38)	<0.0001	0.81 (0.72-0.90)	<0.0001
Country	83.84 (65.83-101.85)	<0.0001	48.26 (39.99-56.53)	<0.0001
Country2	-31.26 (-48.93- -13.49)	0.0786	14.38 (7.32-21.45)	0.0430

* T-test estimate and p-value when Ho: β1 = 1

Table 7: Sub-analysis t-test, ANOVA, and beta estimates for WHO/MOH cases (n=145) and deaths (n=156)

	Cases WHO/MOH (n=145)		**Deaths WHO/MOH (n=156)**	
	Estimate (CI)	**p-value**	**Estimate (CI)**	**p-value**
T-value*	-1.18	0.2399	7.71	<0.0001
F-value	327.00	<0.0001	885.14	<0.0001
R Square	92.2%		95.9%	
Beta Estimates				
Intercept	-47.21 (-59.47- -34.94)	0.0002	-6.42 (10.95- -1.88)	0.1590
News	0.95 (0.90-0.99)	<0.0001	1.17 (1.15-1.19)	<0.0001
Country	27.53 (16.39-38.67)	0.0147	-21.32 (-26.21- -16.44)	<0.0001
Country2	21.91 (11.09-32.73)	0.0449	-2.51 (-7.28-2.25)	0.5988
Referenced data source			13.75 (9.65-17.84)	0.0010
Reporting Location	19.63 (11.33-27.93)	0.0194		
Date	0.41 (0.26-0.57)	0.0092		

* T-test estimate and p-value when Ho: $\beta 1 = 1$

Figures and Figure Legends

Figure 1: Map of 2014 EVD origin Meliandou, Guinea (77)

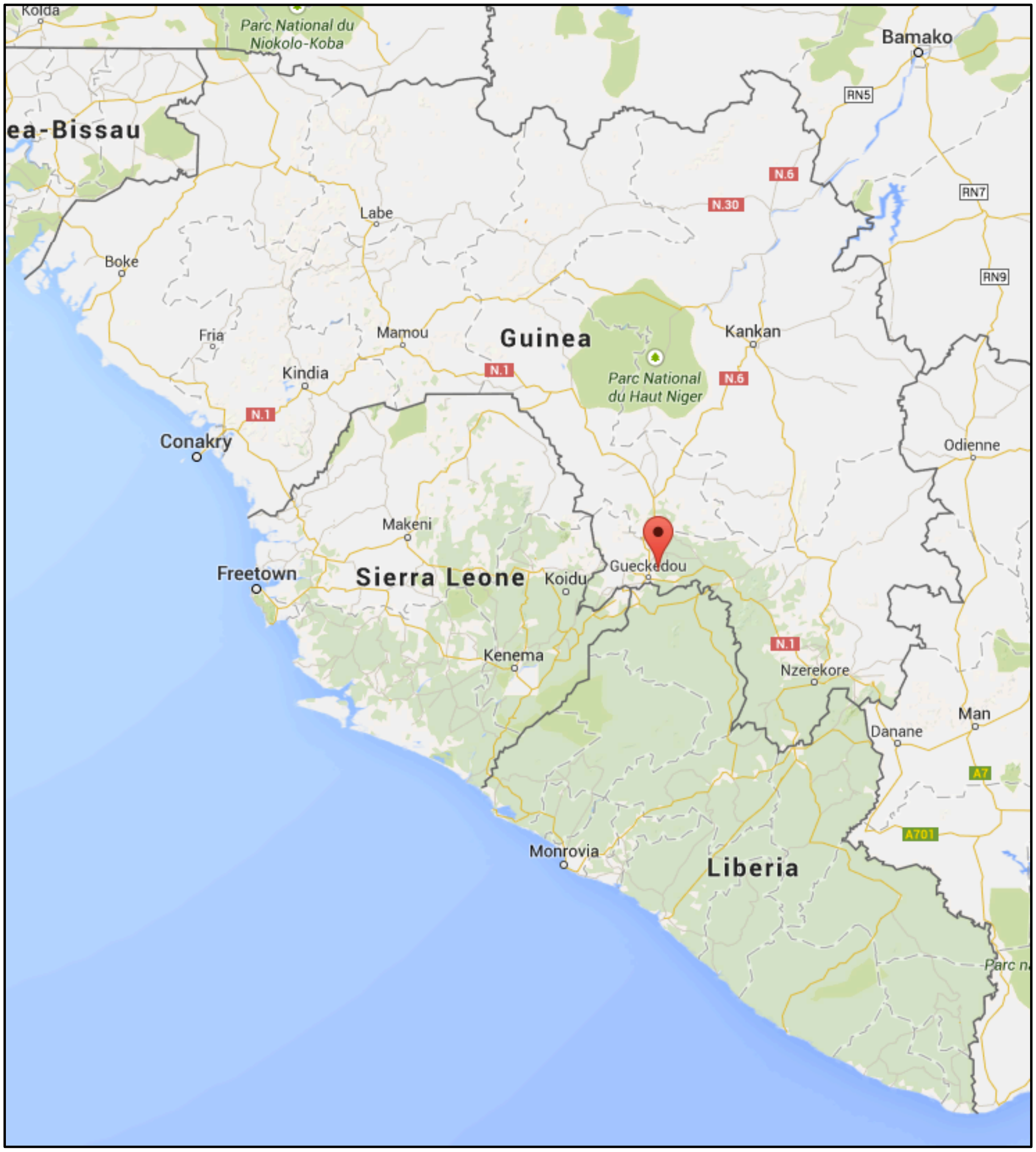
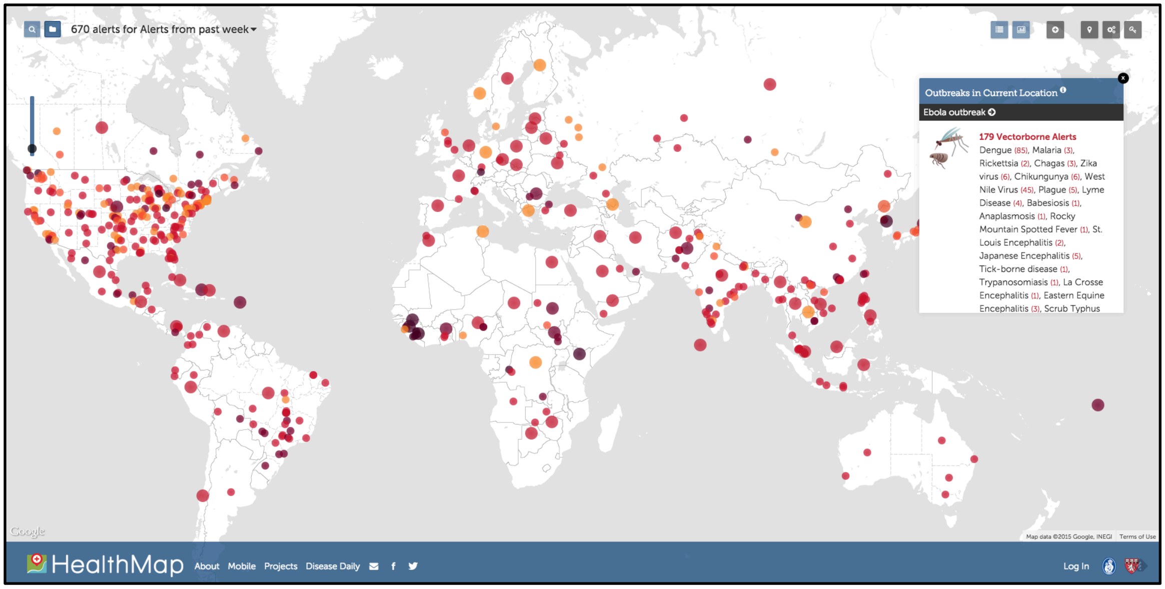


Figure 2: HealthMap: Alerts from past week (July 5, 2015) (78)



670 alerts for Alerts from past week

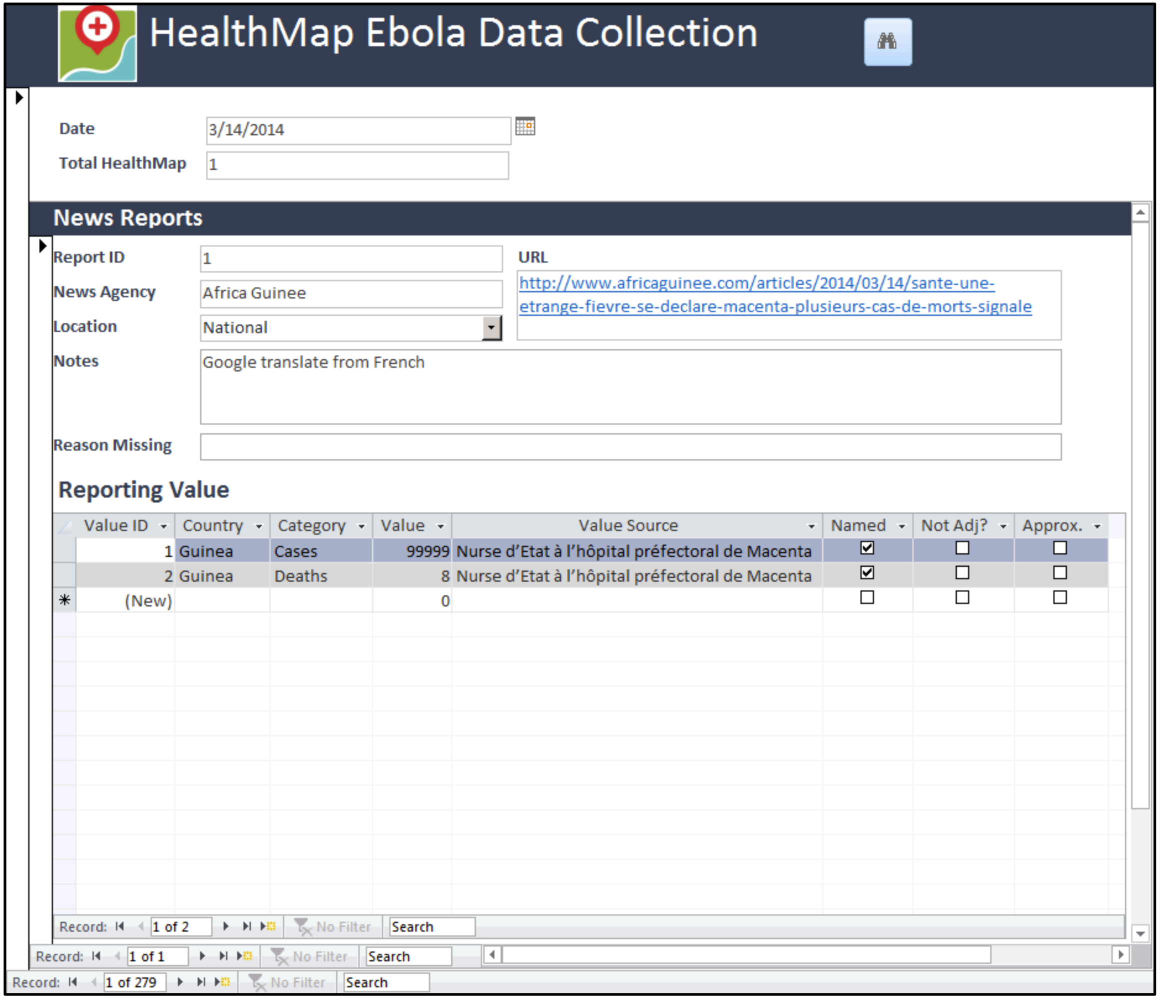
Outbreaks in Current Location

Ebola outbreak

179 Vectorborne Alerts

Dengue (85), Malaria (3), Rickettsia (2), Chagas (3), Zika virus (6), Chikungunya (6), West Nile Virus (45), Plague (5), Lyme Disease (4), Babesiosis (1), Anaplasmosis (1), Rocky Mountain Spotted Fever (1), St. Louis Encephalitis (2), Japanese Encephalitis (5), Tick-borne disease (1), Trypanosomiasis (1), La Crosse Encephalitis (1), Eastern Equine Encephalitis (3), Scrub Typhus

Figure 3: HealthMap Access Data Collection Form from March 14, 2014



HealthMap Ebola Data Collection

Date	3/14/2014
Total HealthMap	1

News Reports

Report ID	1
News Agency	Africa Guinee
Location	National
Notes	Google translate from French
Reason Missing	
URL	http://www.africaguinee.com/articles/2014/03/14/sante-une-etrange-fievre-se-declare-macenta-plusieurs-cas-de-morts-signale

Reporting Value

Value ID	Country	Category	Value	Value Source	Named	Not Adj?	Approx.
1	Guinea	Cases	99999	Nurse d'Etat à l'hôpital préfectoral de Macenta	☑	□	□
2	Guinea	Deaths	8	Nurse d'Etat à l'hôpital préfectoral de Macenta	☑	□	□
(New)			0		□	□	□

Figure 4: Variability between WHO/MOH and news reported case and mortality for Guinea (March 14 – August 28, 2014)

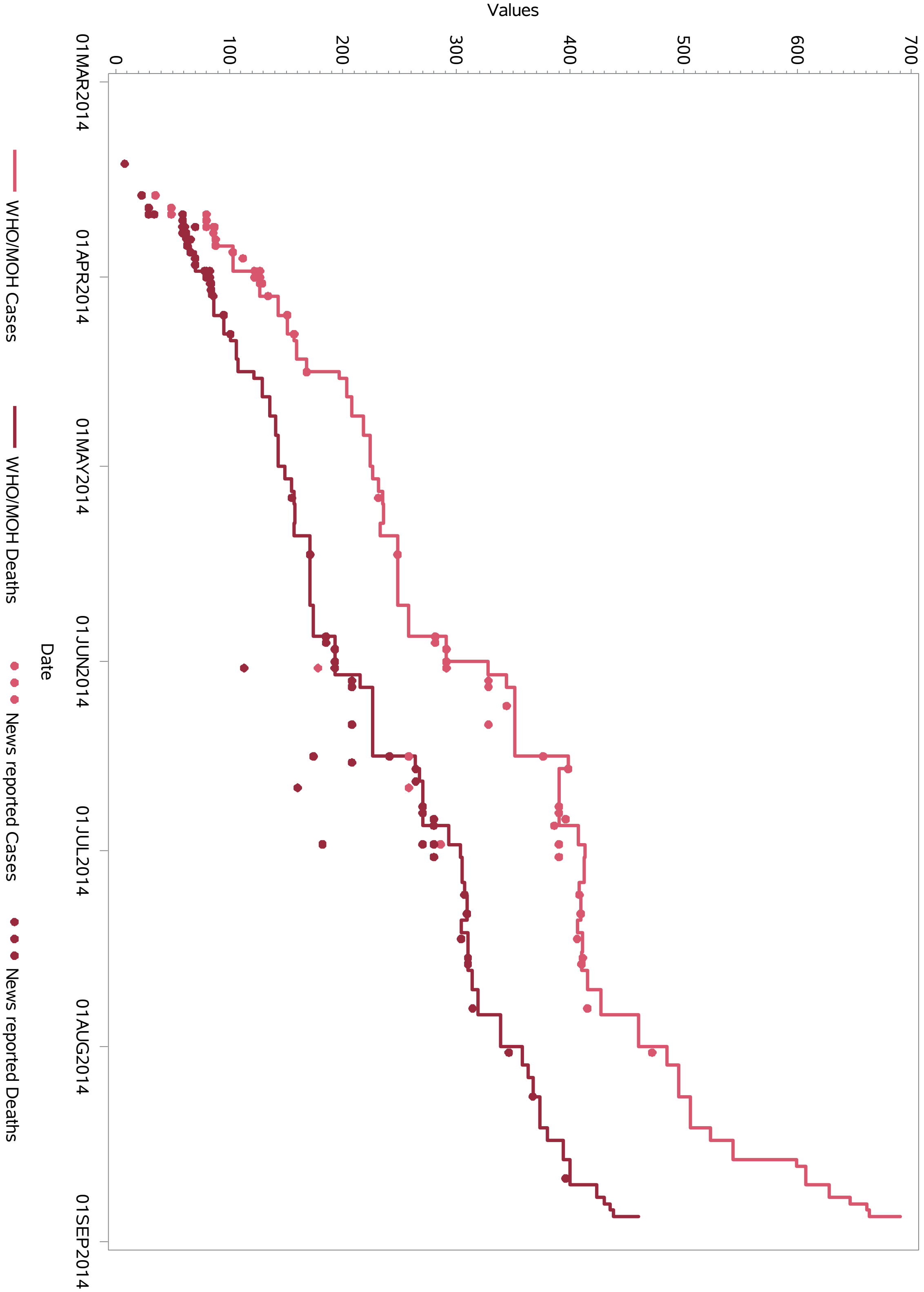


Figure 5: Variability between WHO/MOH and news reported case and mortality for Liberia (March 14 – August 28, 2014)

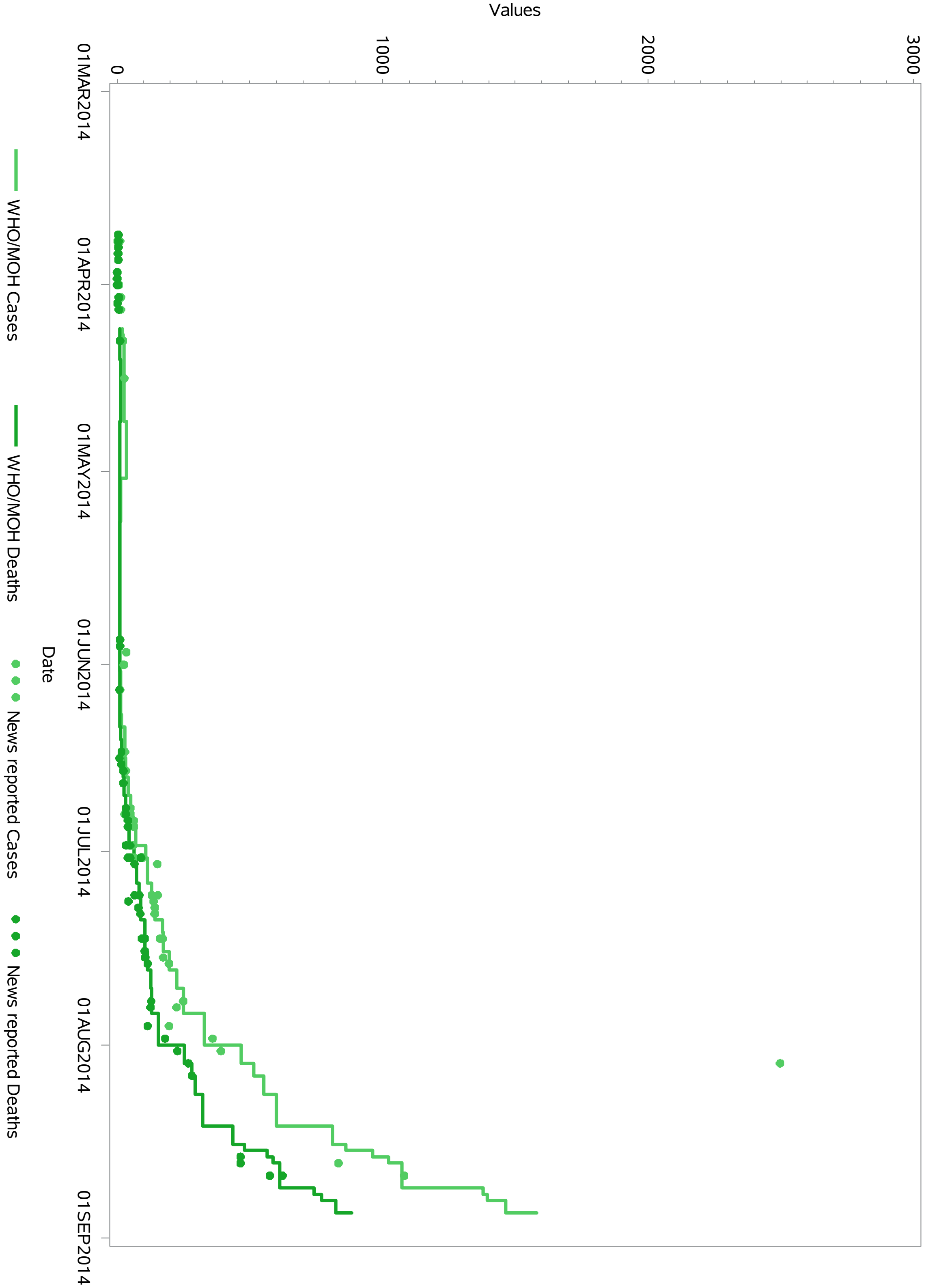


Figure 6: Variability between WHO/MOH and news reported case and mortality for Sierra Leone (March 14 – August 28, 2014)

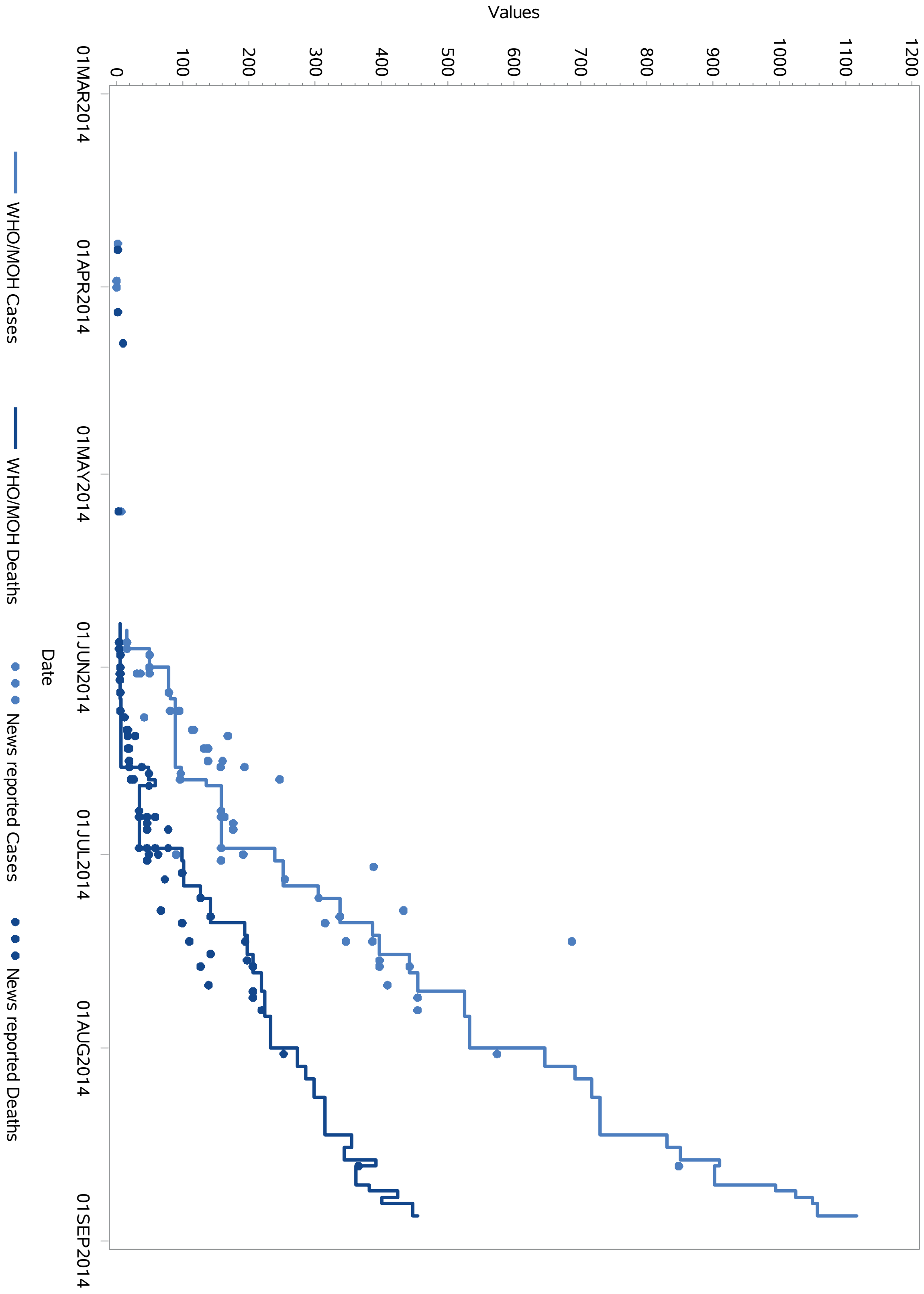


Figure 7: Variability between WHO/MOH and news reported case and mortality for Guinea, Liberia, and Sierra Leone (March 14 – August 28, 2014)

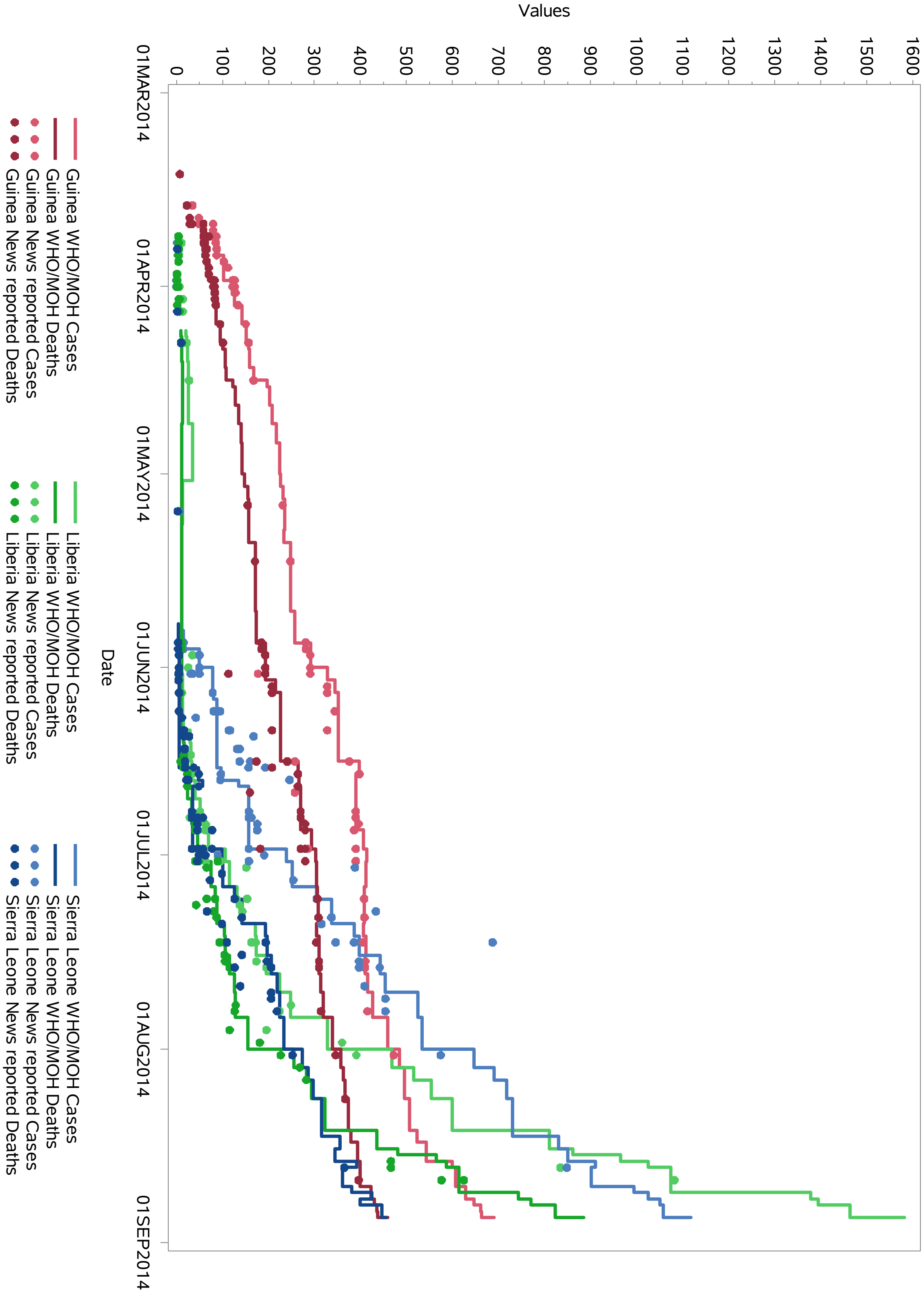
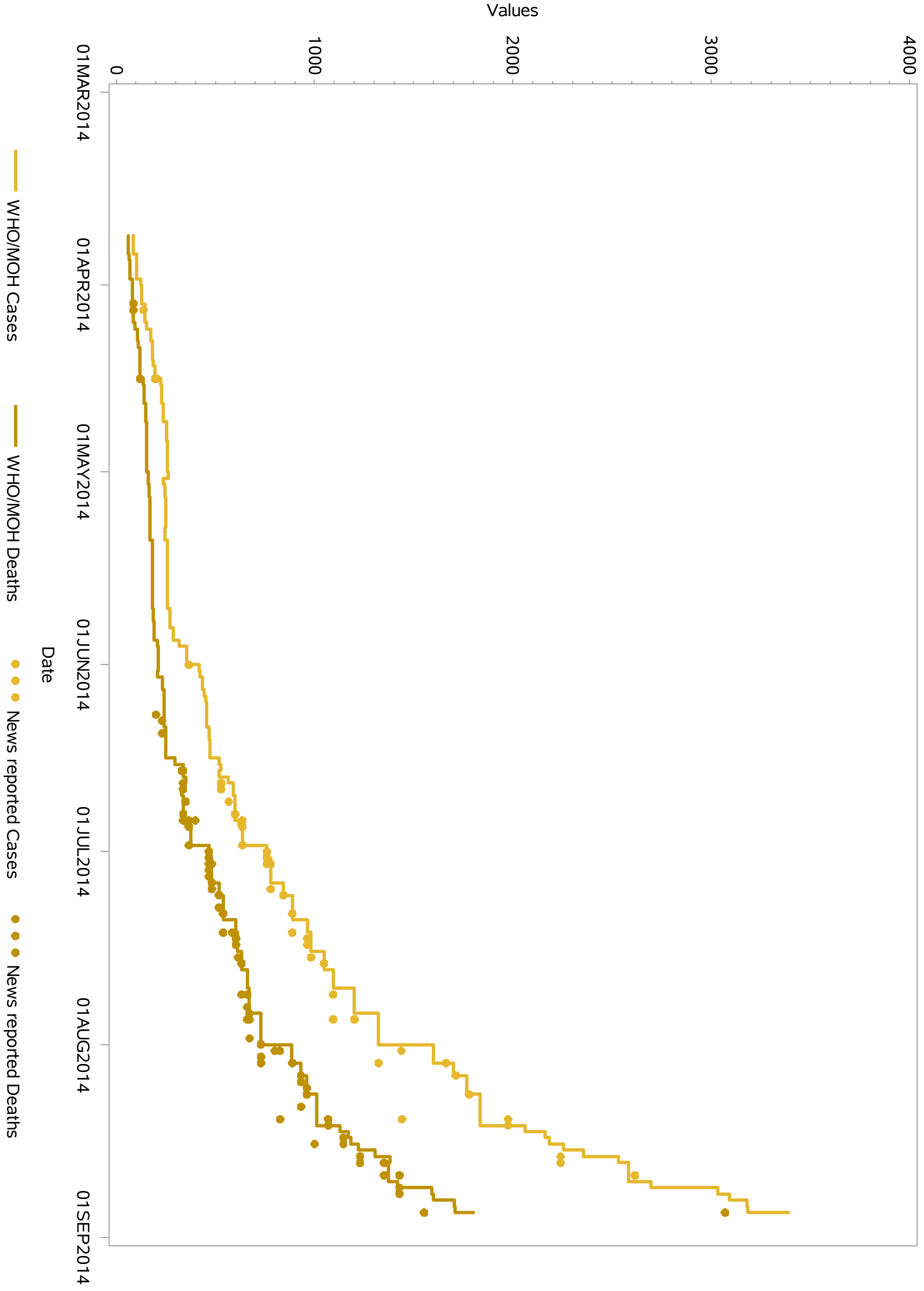


Figure 8: Variability between WHO/MOH and news reported case and mortality for West Africa total (March 14 – August 28, 2014)



Appendices

A. Data Dictionaries

1. Data Dictionary: HDX Sub-national time series (March 23 – August 28, 2015)

Original Data: Humanitarian Data Exchange (HDX): Sub-national time series data on Ebola cases and deaths in Guinea, Liberia, Sierra Leone, Nigeria, Senegal and Mali since March 2014. Collected by OCHA ROWCA, pulled Feb 19, 2015.
Data availability: Data dictionary, cleaned data, and SAS code available via Figshare (79)

Variable:	Format:	Label:	Code:	Notes:
Country	Country3f.	Guinea Liberia Sierra Leone	1 2 3	Additional countries removed; changed coding from categorical to numeric with format[1]
Category	Category3f.	New Cases Cases Deaths	1 2 3	Additional categories removed; changed coding from categorical to numeric with format[2]
Sources	Sources2f.	International National	1 2	Additional categories removed; changed coding from categorical to numeric with format[3]
Value	BEST12.	--	Num	Numeric value representing reported category (ex. 49 “deaths”)
Date	DATE9.	DDMMYYYY	DATE9.	Reported date, not collection date.
Dataset	Datasetf.	HDX HealthMap	0 1	Identify dataset

Localite (105 unique sub-country identifiers) and *Link* (URL) variables deleted from the dataset.

[1] Removed “Nigeria” from dataset (n=95). Remaining variables numerically coded.

[2] Removed “Confirmed Cases” “Probable Cases” and “Suspected Cases” from dataset due to duplication. The frequency of “Cases” (n=741) for the study duration (March 23 - August 28, 2014) within the three countries of interest (excluding Nigeria) did not significantly differ from the summed frequency of “Confirmed Cases” (n=358), “Probable Cases” (n=401) and “Suspected Cases” (n=10); total (n=738). The research chose to keep “Cases” as representative. The remaining categories were numerically coded 1= “New Cases,” 2=“Cases” and 3=“Deaths” as to match the HealthMap data coding.

[3] “GVT”, “Gvt”, and “Ministere de la Sante”, combined into “National” (n= 285) “ECHO”, “UNICEF”, “ WHO”, and “ WHO;gvt” combined into “International” (n=164) in HDX2.

2. Data Dictionary: HealthMap 2014 Ebola Outbreak (March 14–August 28, 2014)

Original Data: data collected and coded by the researcher.
Data availability: Data dictionary, cleaned data, and SAS code available via Figshare (80)

Variable:	Format:	Label:	Code:	Notes:
HMID	--	--	Num	Unique identifier for each HealthMap reporting day
ReportID	--	--	Num	Unique identifier for each HealthMap news report
ValueID	--	--	Num	Unique identifier for each HealthMap counts
Date	DATE9.	DDMMYYYY	DATE9.	Date of HealthMap Report(s)
TotNumHm	--	--	Num	Total number of HealthMap reports on given day
Country	Countryf.	Guinea Liberia Sierra Leone West Africa	1 2 3 4	Lookup Table
Category	Categoryf.	New Cases Cases Deaths New Deaths	1 2 3 4	Lookup Table
Value	--	--	Num	Numeric value representing reported category (ex. 49 “deaths”); 99999=referenced
NewsAgency	--	--	Char	Name of news agency reporting New Cases, Cases, Deaths or New Deaths
NewsAgency_ Collapse	--	--	Char	Name of News Agency collapsing - Reuters, WHO, Bloomberg, and Yahoo!
NewsAgency_ Down	--	--	Char	Name of News Agency keeping all local News Agencies (ie. CTV News)
NewsAgency_ Up	--	--	Char	Name of News Agency keeping all International News Agencies (ie. AP)
Location	Locationf.	Nationally Africa Internationally	1 2 3	Lookup Table - where is the "NewsAgency" located?
Link	--	--	Char	URL
Notes	--	--	Char	Additional Notes

Missing	--	--	Char	Reason for missing category data
NotAdj	Yesnof.	No Yes	0 1	Are the numbers NOT adjusted to only include Guinea, Liberia, and Sierra Leone, yes or no?
Sources_ original	--	--	Char	Name of news agency providing Category values
HM_Sources	Sourcesf.	WHO MOH	0 1	Article source reference for case and mortality counts, either from an international source (WHO) or domestic source (MOH). Dichotomized from Sources_originial.
Named	Yesnof.	No Yes	0 1	Was a contact person associated with ReportSource, yes or no?
Approximate	Yesnof.	No Yes	0 1	Are the values approximate, yes or no?
Dataset	Datasetf.	HDX HealthMap	0 1	Identify dataset
Notes for “Category” qualitative coding: Only stories providing country-level aggregated data was included, no district or sub-district level data was included. Stories breaking the new of person of interest death is not classified as a "new case"				

3. Data Dictionary: WHO/MOH and HealthMap merge

Original Data: Combined HealthMap Ebola March 14 – August 28, 2014 and Humanitarian Data Exchange (HDX): Sub-national time series data on Ebola cases and deaths in Guinea, Liberia, Sierra Leone, Nigeria, Senegal and Mali since March 2014. Collected by OCHA ROWCA, pulled Feb 19, 2015.
Data availability: Data dictionary, cleaned data, and SAS code available via Figshare (81)

Variable:	**Format:**	**Label:**	**Code:**	**Notes:**
HMID	--	--	Num	Unique identifier for each HealthMap reporting day
ReportID	--	--	Num	Unique identifier for each HealthMap news report
ValueID	--	--	Num	Unique identifier for each HealthMap counts
Date	DATE9.	DDMMYYYY	DATE9.	Date of HealthMap Report(s)
Date_count	--	--	Num	Date since January 1, 1960
Date_norm	--	--	Num	Absolutely value of the date since first reporting day in Guinea (34,195), Liberia (34,205), and Sierra Leone (34,207). Used to normalize date.
TotNumHm	--	--	Num	Total number of HealthMap reports on given day
Country	Countryf.	Guinea Liberia Sierra Leone	0 1 2	Country where case or mortality values are coming from
Country1	--	Not Guinea Guinea	0 1	Dummy variable for Country
Country2	--	Not Liberia Liberia	0 1	Dummy variable for Country
NewsAgency	--	--	Char	Name of news agency reporting New Cases, Cases, Deaths or New Deaths
NA	Naf.	Major News Minor News	0 1	Dichotomized *NewsAgency* based on size of news agency
Location	Locationf.	Affected Non-affected	0 1	Reporting from affected country or non-affected country
NotAdj	Yesnof.	No Yes	0 1	Are the numbers NOT adjusted to only include Guinea, Liberia, and Sierra Leone, yes or no?
HM_Sources	Sourcesf.	WHO MOH	0 1	Article source reference for case and mortality counts, either from an international source (WHO) or domestic source (MOH)

Approximate	Yesnof.	No Yes	0 1	Are the values approximate, yes or no?
HM_Cases	--	--	Num	Count for HealthMap cases
HM_Deaths	--	--	Num	Count for HealthMap deaths
HDX_Cases	--	--	Num	Count for HDX cases
HDX_Deaths	--	--	Num	Count for HDX deaths
Log_HM_Case	--	--	Num	Log transformed HealthMap cases
Log_HM_Deaths	--	--	Num	Log transformed HealthMap deaths
Log_HDX_Cases	--	--	Num	Log transformed HDX cases
Log_HDX_Deaths	--	--	Num	Log transformed HDX deaths
Diff_Cases	--	--	Num	Absolute value of the difference between HealthMap cases and HDX cases by country and day
Diff_Deaths	--	--	Num	Absolute value of the difference between HealthMap cases and HDX deaths by country and day

B. Model Selection Process:

1. Model 1 & 2: Full analysis for cases and deaths

Model Selection: Model 1 (Full analysis cases)

	F value	p-value	Intercept	Case News Values	Date	Country	Reference	Reporting Location	News Agency Size	Approximated Values
Forward:	154.98	<0.0001	6.4774	0.3855	2.0641	-41.3656				
Backward:	154.98	<0.0001	6.4774	0.3855	2.0641	-41.3656				
Stepwise:	154.98	<0.0001	6.4774	0.3855	2.0641	-41.3656				

Number of Observations Read (400), Number of Observations Used (178), Number of Observations with Missing Values (222)

Number in Mo	R-Square	Adjusted R-Sq	C(p)	AIC	BIC	MSE	Variables in Model
1	0.5908	0.5885	88.8392	1707.7616	1708.2621	14516.0000	hm_Cases
2	0.6983	0.6949	21.7587	1655.4843	1656.9830	10762.0000	hm_Cases date_norm
3	*0.7277*	0.7230	4.9109	1639.2683	1641.4090	9770.8338	hm_Cases date_norm country*
4	0.7332	0.7270	3.3620	1637.6196	1640.0040	9627.9189	hm_Cases date_norm country NewsAgency_loc
5	0.7338	0.7260	5.0031	1639.2464	1641.7340	9663.6106	hm_Cases date_norm country NewsAgency_loc NA

* Final model: HealthMap Deaths, date (reoriented), and Country

Model Selection: Model 2 (Full analysis deaths)

	F value	p-value	Intercept	Mortality News Values	Date	Country	Reference	Reporting Location	News Agency Size	Approximated Values
Forward:	666.87	<0.0001	3.2639	0.6981	0.7952	-23.8492				
Backward:	666.87	<0.0001	3.2639	0.6981	0.7952	-23.8492				
Stepwise:	666.87	<0.0001	3.2639	0.6981	0.7952	-23.8492				

Number of Observations Read (400), Number of Observations Used (222), Number of Observations with Missing Values (178)

Number in Mo	R-Square	Adjusted R-Sq	C(p)	AIC	BIC	MSE	Variables in Model
1	0.8672	0.8666	75.9181	1711.5879	1712.4894	2210.3079	hm_Deaths
2	0.8868	0.8858	34.5102	1678.1119	1679.4032	1892.4760	hm_Deaths date_norm
3	0.9017	0.9004	3.5318	1648.7756	1650.9391	1650.8796	hm_Deaths date_norm country*
4	0.9024	0.9006	4.1636	1649.3749	1651.6434	1648.0558	hm_Deaths date_norm country Approximate
5	0.9029	0.9006	5.0009	1650.1776	1652.5661	1646.7802	hm_Deaths date_norm country Approximate NewsAgency_loc
6	0.9029	0.9002	7.0000	1652.1766	1654.6303	1654.4327	hm_Deaths date_norm country Approximate NewsAgency_loc NA

* Final model: HealthMap Deaths, date (reoriented), and Country

2. Model 3 & 4: Sub-analysis for cases and deaths

Model Selection: Sub-analysis cases (n=145)

	F value	p-value	Intercept	Case News Values	Date	Country	Reference	Reporting Location	News Agency Size	Approximated Values
Forward:	762.04	<0.0001	-14.1875	1.0199				20.91739		
Backward:	409.55	<0.0001	-16.6913	0.9303	0.4577	-14.3032		19.54851		
Stepwise:	762.04	<0.0001	-14.1875	1.0199				20.91739		
* Model Selection Model DF=3, no dummy variables for Country as in ANOVA										

Number of Observations Read (400) Number of Observations Used (145) Number of Observations with Missing Values (255)

Number in Mode	R-Square	Adjusted R-Sq	C(p)	AIC	BIC	MSE	Variables in Model
1	0.9112	0.9105	16.5541	1137.6494	1139.3180	2520.3696	hm_Cases
2	0.9148	0.9136	12.1701	1133.6518	1135.4035	2435.2762	hm_Cases NewsAgency_loc*
3	0.9181	0.9164	8.1919	1129.8010	1131.7905	2355.5576	hm_Cases date_norm country
4	0.9213	0.9190	4.6439	1126.1518	1128.5324	2281.7312	hm_Cases date_norm country NewsAgency_loc
5	0.9220	0.9192	5.3005	1126.7500	1129.3267	2276.0372	hm_Cases date_norm country NewsAgency_loc NA
6	0.9222	0.9188	7.0000	1128.4347	1131.1397	2287.5497	hm_Cases date_norm country HM_sources NewsAgency_loc NA

Model Selection: Sub-analysis deaths (n=156)

	F value	p-value	Intercept	Mortality News Values	Date	Country	Reference	Reporting Location	News Agency Size	Approximated Values
Forward:	1156.92	<0.0001	-23.92759	1.16		10.57483	13.34364			
Backward:	1156.92	<0.0001	-23.92759	1.16		10.57483	13.34364			
Stepwise:	1156.92	<0.0001	-23.92759	1.16		10.57483	13.34364			
* Model Selection Model DF=3, no dummy variables for Country as in ANOVA										

Number of Observations Read (400) Number of Observations Used (156) Number of Observations with Missing Values (244)

Number in Mode	R-Square	Adjusted R-Sq	C(p)	AIC	BIC	MSE	Variables in Model
1	0.9503	0.9499	34.9118	1013.5601	1014.8439	654.87394	hm_Deaths
2	0.9552	0.9546	18.4704	999.3554	1000.9003	594.11967	hm_Deaths country
3	0.9580	0.9572	9.6397	990.9895	992.9061	559.58216	hm_Deaths country HM_sources*
4	0.9592	0.9581	7.2144	988.5475	990.7287	547.47493	hm_Deaths date_norm HM_sources Approximate
5	0.9601	0.9588	5.9182	987.1548	989.6383	539.26824	hm_Deaths country HM_sources Approximate N
6	0.9606	0.959	6.0452	987.1936	989.9401	536.10501	hm_Deaths date_norm country HM_sources Approximate NA
7	0.9606	0.9587	8	989.1459	992.0049	539.56245	hm_Deaths date_norm country HM_sources Approximate NewsAgency_loc NA