Characterizing Ebola Transmission Patterns Based on Internet News Reports

Julie M. Cleaton,1,2 Cecile Viboud,2 Lone Simonsen,3,4,5 Ana M. Hurtado,2 and Gerardo Chowell1,3

1School of Public Health, Georgia State University, Atlanta2College of Liberal Arts and Sciences, Arizona State University, Tempe3Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, Maryland4Department of Public Health, University of Copenhagen, Denmark5Department of Global Health, Milken Institute School of Public Health, George Washington University, Washington D.C.

Background. Detailed information on patient exposure, contact patterns, and discharge status is rarely available in real time from traditional surveillance systems in the context of an emerging infectious disease outbreak. Here, we validate the systematic collection of Internet news reports to characterize epidemiological patterns of Ebola virus disease (EVD) infections during the West African 2014–2015 outbreak.

Methods. Based on 58 news reports, we analyzed 79 EVD clusters (286 cases) ranging in size from 1 to 33 cases between January 2014 and February 2015 in Guinea, Sierra Leone, and Liberia.

Results. The majority of reported exposures stemmed from contact with family members (57.3%) followed by hospitals (18.2%) and funerals (12.7%). Our data indicate that funeral exposure was significantly more frequent in Sierra Leone (27.3%) followed by Guinea (18.2%) and Liberia (1.8%; χ² test; P < .0001). Funeral exposure was the dominant route of transmission until April 2014 (60%) and was replaced with hospital exposure in June 2014–July 2014 (70%), both of which declined after interventions were put in place. The mean reproduction number of the outbreak was 2.3 (95% confidence interval [CI], 1.8, 2.7). The case fatality rate was estimated at 74.4% (95% CI, 68.3, 79.8).

Conclusions. Overall, our findings based on news reports are in close agreement with those derived from traditional epidemiological surveillance data and with those reported for prior outbreaks. Our findings support the use of real-time information from trustworthy news reports to provide timely estimates of key epidemiological parameters that may be hard to ascertain otherwise.

Keywords. Ebola; transmission patterns; West Africa; hospital transmission; funeral transmission.

The unprecedented Ebola epidemic in West Africa may have been triggered by a single cross-species transmission event from a fruit bat, an event that has been traced back to December 2013 in a forested Guinean region [1]. The Ebola outbreak was not reported to the World Health Organization (WHO) until late March 2014, at which point the outbreak had crossed the porous borders of neighboring Sierra Leone and Liberia [2, 3]. The epidemic gained speed over the summer and started abating as behavioral and medical interventions were put in place in fall 2014. As of 15 July 2015, 27 678 illnesses and 11 276 deaths have been reported to WHO [3], with the overwhelming majority of cases reported in West Africa and likely an underestimate.

The epidemic has tapered off, with low incidence counts reported in the region since April 2015 [3]. A great amount of knowledge on the epidemiology and clinical picture of the disease has been gained (eg, [4–9]). However, the nature of and changes in the characteristics of the contact networks and heterogeneity in risk in various exposure settings remain unclear, with information limited in geographic and temporal scope [10, 11]. Such critical information was not available to inform mathematical models used to project the likely trajectory of the outbreak and the likely benefits of various control interventions.

Here, we contribute to the growing body of digital epidemiology studies that make use of nontraditional online data sources to enhance the detection, forecasting, and response to infectious disease threats [12–15]. Specifically, we show that systematic collection and analysis of Internet news reports from authoritative media outlets and public health authorities can overcome the scarce and patchy information on exposure patterns and chains of transmission available from formal epidemiological surveillance efforts.

MATERIALS AND METHODS

We reviewed news stories and investigative reports published between March 2014 and March 2015 that described suspected, probable, and confirmed cases of Ebola in the 3 most affected countries (Guinea, Sierra Leone, and Liberia) from the WHO website as well as online authoritative media outlets (Supplementary Materials). We screened articles from the Ebola section of news sources or used the search keyword “Ebola” and...
perused the links available in retrieved articles for further information. We selected those articles that contained detailed stories about Ebola case clusters arising within families or through funerals and hospitals. Specifically, we selected 21 articles and 2 videos out of 624 New York Times articles from the section Times Topics, “The Ebola Outbreak in West Africa,” published between 25 March 2014 and 31 March 2015. We also screened 781 articles from the Washington Post published between 22 March 2014 and 31 March 2015 and retrieved articles by searching the keyword “Ebola” anywhere in the article. Of these, we used information from 6 articles to populate our database. We also retrieved useful information from 6 of 211 EbolaDeeply articles published from 15 October 2014 to 31 March 2015 (http://www.eboladeeply.org/). Finally, from the WHO website, we screened 60 articles in the section “Stories from the Field on Ebola” and 96 WHO situational reports published between March 2014 and March 2015. Of these, we retrieved relevant information from 11 articles and 5 WHO situational reports.

Each article was carefully assessed, and demographic, epidemiologic, and clinical information about each case was systematically extracted, including patient age and gender, country where the patient acquired Ebola virus disease (EVD), month and year of symptoms onset, occupational/healthcare worker status, disease outcome (recovery or death), nature of exposure to the Ebola virus (family/household, hospital, funeral, care, sexual, hazardous waste, childbirth, zoonotic), as well as the number of secondary cases generated by the index case in each cluster. If the date of symptoms onset was not given in the article, we used the article’s publication date.

Each Ebola patient was assigned 1 or several types of Ebola exposures based on information provided in the articles, including family/household, hospital, funeral, care, hazardous waste, and sexual, as detailed below. “Family/household” exposure indicates that the family members of the case were ill with Ebola within 3 weeks before his/her symptoms began or that the article stated that the case got sick with Ebola from contact with a family member. Of note, people living in the same household, such as couples living together, were also considered family.

“Hospital” exposure indicates that the case (e.g., patient, visitor, healthcare worker) got infected after spending time in healthcare settings including hospitals, clinics, Ebola treatment units (ETUs), and isolation facilities. “Funeral” exposure indicates that the case was exposed to the Ebola virus while preparing for or attending a funeral and touching the body of an Ebola deceased person within the previous 3 weeks. “Care” exposure indicates that a patient was infected after helping another EVD patient outside of the hospital setting, including transportation to a healthcare setting or attending to an Ebola case. This category excludes healthcare workers exposed in hospital settings but does include family members who were stated as taking care of one another. “Hazardous waste” indicates that the article stated that the case got sick with Ebola after handling or cleaning waste products of an Ebola patient. “Sexual exposure” indicates that the case had not been in direct contact with an active EVD case in the past 3 weeks but had sexual contact with an Ebola recovered case. “Childbirth” indicates that a pregnant woman infected with EVD gave birth and transmitted EVD to the newborn during labor without further contact with the mother or other EVD patients. Patients were considered healthcare workers if they worked in healthcare settings as doctors, nurses, hygienists, lab technicians, hospital administrators, or hospital guards. Cases were considered hospitalized if they were admitted to a hospital, clinic, ETU, or isolation facility. Some patients were exposed in multiple settings, for example, through both care and family; those cases were counted in both categories. Whenever the means of exposure could not be identified from individual articles, we indicated the exposure information as “unknown.” Exposure was indicated as “zoonotic” for the purported index case (single case) of the epidemic in December 2013 [2].

Transmission chains of Ebola infection were explicitly provided in the article or inferred based on chronological information on the timing of symptoms of successive cases available from all but 8 of the selected articles. We also analyzed the number of secondary cases generated by an index case in each Ebola cluster, which was explicitly stated for most clusters, as these were typically limited to family or hospital transmission.

RESULTS

We analyzed 79 EVD clusters (286 cases) that occurred in Guinea, Sierra Leone, and Liberia between January 2014 and February 2015 (Supplementary Table 1). Eight clusters with known exposures originated in Guinea, 21 in Sierra Leone, and 34 in Liberia (Supplementary Table 1). All clusters except 3 involved a single country only. There was a similar fraction of male and female cases (53.7% vs 46.3%), which was consistent across the 3 countries (χ² test; P = .76). The reported cluster size ranged from 1 to 33 cases and included up to 6 disease generations. The mean cluster size was estimated at 4.3 (95% confidence interval [CI], 3.2, 5.3) by fitting a negative binomial distribution to the cluster sizes (Supplementary Figure 1). The curve of monthly cases in our sample was significantly correlated with the outbreak trajectory as reported by WHO (Figure 1; Spearman rho = 0.59, P = .026).

Guinea had the highest rate of cases with unknown exposures (32.3%) followed by Liberia (21.4%) and Sierra Leone (18.5%). The monthly proportion of cases with unknown exposures did not vary significantly during the course of the epidemic (Spearman rho = 0.16, P = .6). Only 5.2% of the cases had multiple exposures, but this proportion did not vary significantly across countries (χ² test; P = .31). The majority of reported exposures stemmed from contact with family members (57.3%) followed by hospitals (18.2%) and funerals (12.7%; Table 1). The proportion of cases reporting family exposures was similar among the
3 countries ($\chi^2$ test; $P = .16$). The overall proportion of family exposure was higher among females compared with males (75.8% vs 54.5%; $\chi^2$ test; $P = .004$). The case fatality rate (CFR) was 74.4% (95% CI, 68.3, 79.8) across all identified cases; there was no difference in CFR between patients reporting family exposure and nonfamily exposure ($\chi^2$ test; $P = .46$) or among countries (Guinea, 72.4%; Sierra Leone, 72.7%; Liberia, 75.8%; $\chi^2$ test; $P = .87$).

The proportion of hospital exposures varied significantly across the 3 countries, with the highest proportion reported in Liberia (24.5%) followed by Sierra Leone (13.6%) and Guinea (9.1%; $\chi^2$ test; $P < .04$). A greater fraction of hospital-based transmission involved males compared with females ($\chi^2$ test; $P < .0001$). We did not find a significantly different CFR between individuals with and without hospital exposure ($\chi^2$ test; $P = .503$).

Figure 1. Monthly numbers of Ebola virus disease (EVD) cases based on case reports of the World Health Organization (WHO) in Guinea, Sierra Leone, and Liberia and our sample of cases clusters, January 2014–February 2015.

Table 1. Ebola Exposure Patterns by Country, Gender, Disease Outcome, and Healthcare Worker Status, December 2013–February 2015

<table>
<thead>
<tr>
<th>Exposure Type, %</th>
<th>All</th>
<th>Country</th>
<th>Gender</th>
<th>Disease Outcome</th>
<th>Healthcare Worker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Guinea</td>
<td>Sierra Leone</td>
<td>Liberia</td>
<td>Male</td>
</tr>
<tr>
<td>Family/household</td>
<td>44.1</td>
<td>33.8</td>
<td>42.0</td>
<td>50.0</td>
<td>40.0</td>
</tr>
<tr>
<td>Hospital</td>
<td>14.0</td>
<td>6.2</td>
<td>11.1</td>
<td>19.3</td>
<td>18.1</td>
</tr>
<tr>
<td>Funeral</td>
<td>9.8</td>
<td>12.3</td>
<td>22.2</td>
<td>1.4</td>
<td>7.6</td>
</tr>
<tr>
<td>Care</td>
<td>5.2</td>
<td>9.2</td>
<td>1.2</td>
<td>5.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Sexual</td>
<td>1.7</td>
<td>0</td>
<td>4.9</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>Hazardous waste</td>
<td>1.4</td>
<td>4.6</td>
<td>0</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>Childbirth</td>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>Zoonotic</td>
<td>0.3</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Unknown</td>
<td>23.1</td>
<td>32.3</td>
<td>18.5</td>
<td>21.4</td>
<td>26.7</td>
</tr>
<tr>
<td>Multiple exposures</td>
<td>5.2</td>
<td>4.6</td>
<td>2.5</td>
<td>7.1</td>
<td>2.9</td>
</tr>
</tbody>
</table>

In the case of multiple exposures, multiple exposure types were counted.
*Care exposure indicates that a patient was infected after helping another Ebola virus disease patient outside of the hospital setting, including through transportation to a healthcare setting or attending to a patient case.
Funeral exposure was significantly more frequent in Sierra Leone (27.3%) followed by Guinea (18.2%) and Liberia (1.8%; $\chi^2$ test; $P < .0001$). Nine clusters in our data included funeral transmission; in 6 of these 9, funeral transmission was the initial exposure event. Moreover, we did not find a significant difference in the proportion of funeral exposures by gender ($\chi^2$ test; $P < .133$).

Overall, the proportion of hospital exposures peaked in April 2014 (70%) and declined to low levels during the subsequent

---

**Figure 2.** Temporal variation in the distribution of Ebola exposures through family, hospitals, and funerals in Guinea, Sierra Leone, and Liberia, January 2014–February 2015.

**Figure 3.** Temporal variation in the frequency of healthcare worker cases among all cases in our sample based on case clusters retrieved from published literature including newspaper reports compared with the proportion of healthcare worker cases from all Ebola cases reported by the World Health Organization (WHO), January 2014–February 2015.

---
epidemic months (Figure 2). Similarly, the proportion of funeral exposures was highest (60%) during June 2014–July 2014 and then declined during the later months of the epidemic (Figure 2). By August 2014, the family/household became the dominant setting of exposure, accounting for 60% of new cases.

**Figure 4.** Distribution of the reproduction number estimates for index cases in our sample of clusters, January 2014–February 2015.

**Figure 5.** Temporal distribution of the reproduction number estimates available from clusters, January 2014–February 2015. The horizontal dashed line at \( R = 1 \) is shown for reference.
Our data included 46 cases who were healthcare workers (16.1%). Of note, only 7 healthcare workers in 5 clusters transmitted EVD to 15 individuals: 9 through family exposure, 2 through care outside of the hospital, and 4 through hospital exposure. The proportion of healthcare worker cases peaked in June 2014 and declined over the subsequent months of the epidemic (Figure 3). Moreover, healthcare workers were more likely to be admitted to hospitals or ETUs compared with non-healthcare workers (91.7% vs 59.6%; χ² test; P = .003).

Based on the identified chains of transmission, we estimated 39 individual level reproduction numbers, ranging from 1 to 7. By fitting a negative binomial distribution to our set of reproduction numbers, we estimated the mean reproduction number at 2.3 (95% CI, 1.8, 2.7; Figure 4). Estimates of the reproduction number did not vary significantly across countries (analysis of variance test; P = .75). The mean R from data during January 2014–July 2014 (R = 2.8; 95% CI, 1.2, 4.4) was not significantly different from the mean R estimated based on the remainder of the epidemic during August 2014–February 2015 (R = 2.1; 95% CI, 1.7, 2.5; Figure 5).

We did not find a significant change in the monthly estimates of the CFR over the course of the epidemic (χ² test; P = .09). Similarly, the age-specific CFR appeared higher (70% and 82%) among those aged <15 years and >45 years compared with 56% among individuals aged 15–44 years (Supplementary Figure 2). However, this age difference did not reach statistical significance (χ² test; P = .15) as our sample was reduced to a small fraction of cases with available age information (29%).

**DISCUSSION**

Here, we have harnessed the potential of real-time publicly available Internet reports from the WHO website as well as authoritative media outlets to identify key information on exposure and transmission patterns during the 2014–2015 West African Ebola epidemic. Our findings on exposure patterns are in good agreement with those derived from epidemiological surveillance data [10] and with those reported for prior outbreaks [1–3,4], highlighting the predominance of hospital and funeral transmission in the pre-intervention period. Similarly, trends in the frequency of hospital exposures aligned well with the proportion of patients among healthcare workers from official WHO statistics. Our mean estimate of the reproduction number based on cluster data lies toward the higher end of the range of previously published estimates, that is, around 1.5–2.5 [7,8,16–18]. This is likely explained as a reporting bias, since news reports favor larger clusters.

Faye et al [10] analyzed a transmission tree composed of 152 Ebola cases from 3 areas of Guinea between February 2014 and August 2014. The relative exposure patterns for Guinea in our data are consistent with those reported by Faye et al [10]. However, our dataset included cases up to February 2015 [10]. Specifically, Faye et al found a predominance of community exposure (72% through family), followed by 13% through hospital exposure and 13% through funeral exposure [10]. In comparison, our Guinean data indicated that 50% of cases were exposed through family, 9.1% via hospitals, and 18.2% via funerals. Moreover, Faye et al [10] reported the proportion of healthcare workers in the transmission tree at 14%, which is similar to the proportion in our data for Guinea (16.7%). The hospitalization rate in our data was also similar to that reported by Faye et al (80% vs 81%) [10].

Overall, in our data, exposure via family contacts (57%) was the most frequent during the epidemic, which is in line with exposure patterns from prior Ebola outbreaks [19–22] and with chains of transmission for the ongoing epidemic in Guinea (February 2014–August 2014) [10]. In contrast, a limited outbreak of EVD in Nigeria that stemmed from a single case importation that originated from Liberia was amplified by hospital exposures that primarily affected healthcare workers (11/20) [11,23].

The proportion of funeral and hospital exposures in our data peaked in April 2014 and June 2014, respectively, which is a signal of super-spreading events that facilitated rapid spread of the virus across communities. Hospital exposures declined considerably after July 2014, likely as a result of the improvement in infection control measures in healthcare settings. Similarly, funeral exposures occurred sporadically in later epidemic months, which suggests the positive effect of massive educational campaigns undertaken in the region. The decline in hospital-based transmission is in line with a decline in the proportion of healthcare workers in our data and in statistics retrieved from the WHO situational reports [24]. Due to our small sample size, it was not possible to characterize temporal changes in exposure patterns by country; however, temporal trends in frequency of nosocomial and funeral exposure reported in news articles are a good marker of the effectiveness of interventions.

Sexual transmission appears to have been infrequent during the epidemic, associated with 2.0% of known exposures in our data. However, this transmission pathway remains a concern, even as the epidemic is waning, because of its potential to generate flare-ups of the disease, even after active Ebola transmission chains have been halted. Continued vigilance and monitoring of the populations months after the end of the epidemic will be critical to ensure that disease flare-ups are quickly contained.

Our mean estimate of the reproduction number at 2.3 lies in the higher end of published estimates, including estimates for past outbreaks in Central Africa [25,26] and estimates derived for the 2014–2015 Ebola epidemic using incidence time series in West Africa [7,8,16–18] or a transmission tree of Ebola cases in Guinea during March 2014–August 2014 [10]. Our estimate is also in agreement with those reported for prior Ebola outbreaks [25,26]. Nevertheless, individual level estimates of the reproduction number for Ebola show substantial heterogeneity (range 1–7 in our data), which underscores the potential role of super-spreading events in sustaining local transmission and a potential opportunity to guide intervention strategies [27].
Moreover, it is worth noting that the reproduction number of the Ebola epidemic at the district level tends to quickly decline after just a few generations of infections. This could result from the natural history of disease transmission in a highly clustered network [28], or it is possible that the benefits of control measures/interventions combined with behavioral changes [6]. We were, however, not able to characterize a particular temporal pattern in the effective reproduction number owing to our small sample size.

The higher proportion of cases reported for Liberia (49%) relative to Sierra Leone (28.3%) and Guinea (22.7%) in our news sample likely reflects an American-centric bias toward reporting cases from a country with strong ties to the United States. Moreover, age information was missing in 71% of the cases, hampering our ability to analyze the transmission patterns according to age. Further, compared with official WHO statistics (3.5%), our sample contained a higher proportion of healthcare workers (16.1%). Other potential biases might exist regarding the writing of survivor stories or sensational articles. However, it is reassuring that our regional estimate of the CFR at 74.4% and associated age patterns are in good agreement with those reported by the WHO Ebola response team [4, 29]. Finally, we focused our search on 2 major newspapers, an Ebola-specific website, and WHO reports so as to maximize the amount of detailed information on Ebola clusters provided by original news stories. Other media sources such as Twitter and Promed have also provided useful epidemiological data [12, 30] in past outbreaks. However, in our experience, these media did not focus on detailed stories of Ebola transmission within individual clusters. Casting a broader net would have probably resulted in a larger sample size but would have required systematic text processing to discard noise from relevant information.

Only a few studies have assessed the role of real-time nontraditional online data sources in elucidating epidemiological characteristics of epidemic outbreaks. For instance, disease outbreaks captured from Internet data collected by HealthMap [31] have been shown to provide epidemiological information ahead of formal epidemiological surveillance reports, including during the early phase of the 2010 Haitian cholera outbreak [32, 33]. More recently, an association between the reproduction number and the volume of news reports documenting public health interventions and aggravating events relating to the Ebola epidemic in Sierra Leone and Liberia was reported [34]. Here we have provided further evidence that information derived from the systematic collection and analysis of data from unstructured news reports from authoritative sources may provide a reliable means to assess epidemiological patterns of evolving infectious disease threats in near real time.

Supplementary Data

Supplementary materials are available at http://cid.oxfordjournals.org. Consisting of data provided by the author to benefit the reader, the posted materials are not copyrighted and are the sole responsibility of the author, so questions or comments should be addressed to the author.

Notes

Financial support. C. V. and G. C. acknowledge financial support from the Division of International Epidemiology and Population Studies, the Fogarty International Center, US National Institutes of Health (NIH), funded in part by the Office of Pandemics and Emerging Threats at the US Department of Health and Human Services. G. C. also acknowledges support from the National Science Foundation (NSF; grant number 1414374) as part of the joint NSF–NIH–US Department of Agriculture Ecology and Evolution of Infectious Diseases Program, United Kingdom Biotechnology and Biological Sciences Research Council (grant number BB/M008894/1; RAPIDD NSF grant number 1518939; and NSF grant number 1518529; and NSF-Division of Information & Intelligent Systems grant number1518939). L. S., C. V., and G. C. acknowledge support from the Research and Policy for Infectious Disease Dynamics Program of the US Department of Homeland Security, and L. S. received support from the Lundbeck Foundation, Denmark.

Potential conflicts of interest. All authors: No potential conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References


