Epidemiol. Infect. (2017), **145**, 1993–2002. © Cambridge University Press 2017
doi:10.1017/S0950268817000760

Constructing Ebola transmission chains from West Africa and estimating model parameters using internet sources

W. B. P. PETTEY[1,2], M. E. CARTER[1,2], D. J. A TOTH[1,3], M. H. SAMORE[1,2] AND A. V. GUNDLAPALLI[1,2]*

[1] *University of Utah School of Medicine, Salt Lake City, Utah, USA*
[2] *VA Salt Lake City Health Care System, Salt Lake City, Utah, USA*
[3] *University of Utah, Salt Lake City, Utah, USA*

Received 25 March 2016; Final revision 10 February 2017; Accepted 24 March 2017; first published online 2 May 2017

SUMMARY

During the recent Ebola crisis in West Africa, individual person-level details of disease onset, transmissions, and outcomes such as survival or death were reported in online news media. We set out to document disease transmission chains for Ebola, with the goal of generating a timely account that could be used for surveillance, mathematical modeling, and public health decision-making. By accessing public web pages only, such as locally produced newspapers and blogs, we created a transmission chain involving two Ebola clusters in West Africa that compared favorably with other published transmission chains, and derived parameters for a mathematical model of Ebola disease transmission that were not statistically different from those derived from published sources. We present a protocol for responsibly gleaning epidemiological facts, transmission model parameters, and useful details from affected communities using mostly indigenously produced sources. After comparing our transmission parameters to published parameters, we discuss additional benefits of our method, such as gaining practical information about the affected community, its infrastructure, politics, and culture. We also briefly compare our method to similar efforts that used mostly non-indigenous online sources to generate epidemiological information.

Key words: Disease transmission modeling, Ebola virus.

INTRODUCTION

The Ebola virus disease epidemic in 2014 highlighted barriers to timely, accurate, sufficiently detailed, and accessible case and community data. Mathematical modelers of infectious disease need parameters derived by reported transmission events, including who infected whom and when, to generate appropriate model transmission parameters. In addition, they need to know the nature of and context surrounding contacts, both of which come into play when trying to forecast disease. Data meeting these requirements for the Ebola virus disease epidemic were not available. In spite of best efforts at forecasting through a very engaged modeling community, the paucity of case transmission data was cited as one of the reasons for disease forecasts that greatly overestimated the

* Author for correspondence: A. V. Gundlapalli, MD, PhD, MS, Associate Professor, Departments of Internal Medicine, Pathology, and Biomedical Informatics, University of Utah School of Medicine, 30 N. 1900 E. Room 5B114D SOM, Salt Lake City, Utah 84132, USA.
(Email: adi.gundlapalli@hsc.utah.edu)
Article Summary Line: We constructed an infectious disease transmission chain using only publicly available internet sources, which can offer a timely snapshot of what is happening in the community and has the potential to yield detailed transmission trees and estimates for epidemiological parameters.

number of likely cases [1]. The total outbreak size is now estimated at <29 000 documented cases [2], and yet early model projections forecasted more than 1 million cases.

Late 2014 saw many online Ebola virus disease transmission stories events recounted online, some with the level of detail necessary for generating model parameters. Some of these online transmission accounts included remarkable detail of transmission events, including dates of contact, contact names, and the dates of symptom onset. The distribution of individual incubation periods (the period between infection and clinical presentation of symptoms) is important for assessing the duration of quarantine strategies [3]. The distribution of serial intervals or generation times between subsequent cases in transmission chains is required to extract estimates of the population reproductive number from time series of case counts [4] and is useful for assessing the effectiveness of patient isolation in preventing transmission during various stages of disease progression [5]. Finally, the variability in the number of transmissions from individual cases is useful for characterizing new outbreak probabilities and the circumstances of potential superspreading events [6].

We set out to discover whether, using only online publicly available sources, we could discover sufficient details to create an accurate Ebola virus transmission chain for cases that were diagnosed and documented during May–October 2014 and thereafter use this transmission tree to generate reliable estimates for key disease parameters. We present our method for building a transmission chain, briefly compare it to published examples for two clusters, examine the reliability of the associated disease transmission parameters, and then discuss the benefits and challenges of our approach. We conclude with thoughts on generalizability of this approach and application to the next public health emergency.

METHODS

Between August and October 2014, we conducted a human search and review of publicly available Internet resources to find and record person-level accounts of Ebola transmission in West Africa, especially focusing on news stories and online sources originating from the affected countries. Our preferred (sought out) online sources were those that had records of reporting as news organizations for the populations in the affected areas. Online, hyperlinked lists of newspapers by country provided one means for identifying these news sources. Through the advanced search features in popular internet search engines (including Google Search), we were able to conduct internet searches with query returns as they appeared in Guinea, Liberia, Nigeria, and Sierra Leone. (Most news sources identify their geographic distribution and scope by regions, communities, or countries, though we also checked the masthead (or its digital equivalent) to verify where the sources were based.) We adopted the heuristic that first-hand, primary accounts of people, places, and transmission events were reliable starting points for building the transmission narrative. In particular, we identified text within these stories that detailed who acquired infection from whom, dates, locations, and details of symptom onset, quarantine, isolation, resolution, and exposure, if known. We grouped sources by the events they relayed, comparing accounts and attempting to corroborate details between stories and checking that they made sense from an epidemiological standpoint. We continued building the chain until we could no longer find earlier cases and the most recently exposed people were at that time still in their incubation periods.

Curating the transmission chain often required matching imprecisely stated event dates, places, and other details. We recorded vague or imprecise details such as *sometime during the previous week* or *during her hospitalization* with ranges of dates or simple statements of fact (e.g., *Case A and Case B were quarantined on the same day*). Documenting details that were discovered during our search on a desktop calendar proved to be helpful, as it was common to report events according to days of the week, such as *last Thursday* or *the prior weekend*. We also had to be cognizant of times and dates, as, for example, 'now' in Utah, USA is different than 'now' in Lagos, Nigeria. Discrepancies between and across sources were resolved using our best judgement, based on subject matter expertise in infectious diseases, including Ebola, while others that could not be precisely reconciled were recorded and reported as uncertainties.

We fit the gamma distribution to these data on incubation period and serial interval we collected. The gamma distribution was chosen for consistency in comparing our results to those derived from World Health Organization (WHO) data [5]. For the incubation period, we established a date or range of dates of exposure and symptom onset for $N = 23$ individuals. We first treated this information as doubly interval-censored

https://doi.org/10.1017/S0950268817000760 Published online by Cambridge University Press

data [7], meaning that both the time of exposure and time of symptom onset were not precisely known but fell within finite intervals. Even when a precise date for an event was reported, we assumed an interval for the precise timing of the event across the 24-h period of that date. We then reduced the data to single intervals for each individual $i = 1,\ldots,N$, representing possible incubation periods, where the minimum possible incubation period $T^i_{\min}$ was the minimum symptom onset time minus the maximum time of exposure, and the maximum possible incubation period $T^i_{\max}$ was the maximum symptom onset time minus the minimum exposure time. We then optimized the gamma distribution parameters by maximizing the likelihood function

$$L(\theta) = \prod_{i=1}^{N} \{F_\theta(T^i_{\max}) - F_\theta(T^i_{\min})\},$$

where F_θ is the cumulative distribution function of the gamma distribution with parameter set θ. Confidence intervals were constructed using 10 000 bootstrap resamples of the dataset. We found that the more complicated likelihood function acting on the full doubly interval-censored data produced nearly the same maximum likelihood estimate, so we chose to use the simpler likelihood function above to reduce the computational time required to perform sufficient bootstrapping.

For the serial interval, we collected the range of possible symptom onset times for $N = 28$ pairs of Ebola patients in which one patient was identified as the source of infection of the second patient in each pair. We then used the same likelihood function above, where $T^i_{\min}$ and $T^i_{\max}$ were defined as the minimum and maximum possible intervals, for patient pair i, between symptom onset times of the index patient and the patient who acquired infection from that index patient.

This study was reviewed by the Institutional Review Board (IRB) of the University of Utah School of Medicine and was determined to be exempt from IRB oversight.

RESULTS

We accessed a total of 5340 web pages from 293 unique web domains (example: liberianobserver.com) between August 1 and October 31, 2014 in our search for information to build an Ebola transmission chain for cases reported in West Africa from May 1 to October 31, 2014. The time invested totaled approximately 60 h, with half of these hours focused on the Nigeria and St Joseph's Catholic Hospital (Monrovia, Liberia) clusters.

From our internet search, a final set of 116 online news stories were used to build the transmission chain segments. We focused especially on the segment representing transmission in the Nigeria cluster and St Joseph's Catholic Hospital cluster, which is displayed as Figure 1. We tried, as often as possible, to use indigenously produced news and web sites from the affected countries. Sources included, but were not limited to, news articles from Nigeria, including *Vanguard*, *Punch*, *Premium Times*, *Guardian*, *Observer*, and from Liberia, including *FrontPage Africa*, *Liberian Observer*, *Liberian Times*, and *The Inquirer* [7–11]. (A list of these online source links and which ones we used to build Figure 1 is available in the Supplementary Material.) We used blogs, news aggregation services, and televised accounts of transmission as well, but generally these served to improve our search queries. The social networking site Facebook was not a primary source, but was helpful for verifying dates, especially when connected friends and family wrote about loved ones who were sick or deceased. Each news source had the potential to confirm or refute facts, or to open additional paths for investigation through new clues, often by offering a different spelling for names, places, and events. To provide further context to our search strategy, Table 1 provides a classification of these 116 online sources by type of epidemiological information gleaned and by origin of the online source.

The transmission chains we constructed included three-related segments comprised of 59 symptomatic individuals who were infected with Ebola virus between May and September 2014 in Guinea, Sierra Leone, Liberia, and Nigeria. (For comparison, contact tracing efforts in CDC reports that contact tracing efforts in Nigeria alone identified 894 contacts, requiring an estimated 18 500 face-to-face visits [12].) Three major sections of the chain included Ebola transmission involving an herbalist who seeded one of the major Sierra Leone chains, a chain arising in Redemption Hospital in New Kru Town, Monrovia, Liberia, and a chain comprised of two large clusters (St Joseph's Catholic Hospital in Monrovia, Liberia, and in Nigeria) stemming from a single Ebola case (the left-most line in Fig. 1) leading to 37–38 infections in this section. For this paper, we focused our efforts on refining the transmission chain from the two clusters in Nigeria and at St Joseph's Catholic Hospital (illustrated in Fig. 1), and the transmission parameters we present are based on those two clusters

https://doi.org/10.1017/S0950268817000760 Published online by Cambridge University Press

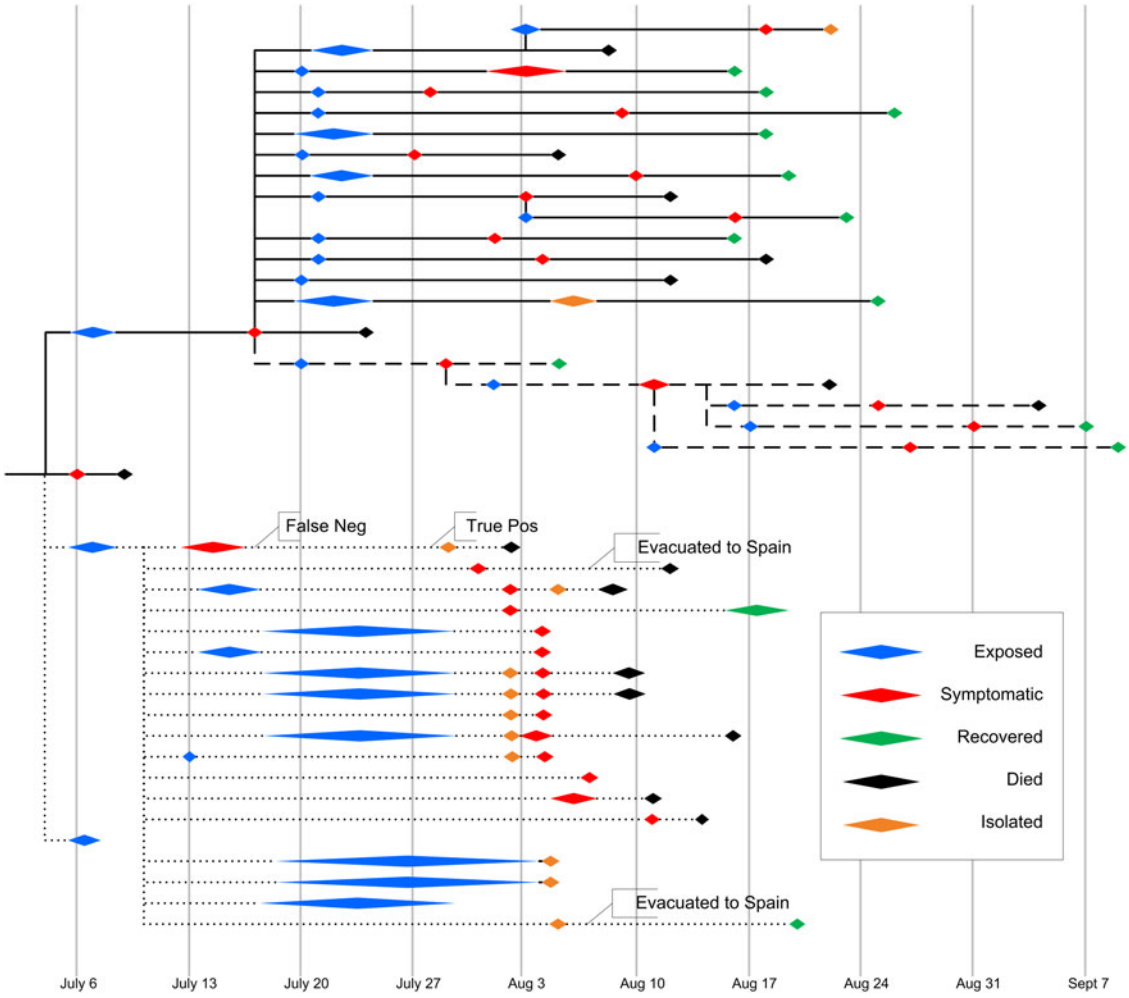

Fig. 1. Timeline showing transmission of Ebola virus disease developed from online, publicly available sources. Diamonds represent noteworthy developments in Ebola infections. They are centered on the dates we identified, and elongated diamonds represent uncertainty in the dates (multiple exposures may have occurred during some of these periods). Solid lines represent transmissions at First Consultants Hospital in Lagos, Nigeria. Dashed lines represent transmissions in Port Harcourt, Nigeria. Dotted lines represent transmission at St Joseph's Catholic Hospital in Monrovia, Liberia. All cases in these clusters originated with a single case, represented with the line entering the figure from the far left-hand side. This same figure, annotated with the online sources, is available in the Supplementary Material.

(we included the online sources for all three segments in the Supplementary Material). Compared with published sources such as Morbidity and Mortality Weekly Report [12] and WHO 2014 reports (published in early and late October, 2014, respectively), we estimate that a low-resolution chain could have been constructed by mid-September 2014, but the final resolution with reliable parameters would have required a further 3–4 weeks (mid-to-late October) for the information to be available from the online sources we used.

Based on the maximum likelihood fit of the gamma distribution, we estimated the mean incubation period to be 12·5 days (95% CI 10·6–14·5 days). The 5th percentile incubation period result was 6·3 days (4·8–8·8 days) and the 95th percentile was 20·4 days (16·5–23·5 days). The mean serial interval was 19·4 days (17·6–21·3 days) with standard deviation 5·1 days. In Table 2, we compare our mean incubation and serial interval estimates to other published results, including those reviewed in Van Kerkhove *et al*. Our range for the mean incubation period overlaps the 9–12 day range of estimates from these other studies, but our mean serial interval estimate was higher (Table 2). Our result of 20·4 days for the 95th percentile incubation period was quite similar to the WHO Ebola Response Team's estimate of 21 days [5] in 2014, which at the time was the first evidence that

https://doi.org/10.1017/S0950268817000760 Published online by Cambridge University Press

Table 1. *Types of online sources used to build the Figure 1 transmission chain by type of online source and type of information helpful for building a transmission chain*

	Origin of online source type				
	State $n = 11$	Country $n = 72$	International $n = 24$	Cannot classify $n = 9$	Total $n = 116$
Epidemiologically important information from source (n, % origin of online source type)					
Person	9 (0·82)	69 (0·96)	20 (0·83)	4 (0·44)	102 (0·88)
Case history	7 (0·63)	46 (0·64)	12 (0·5)	3 (0·33)	68 (0·59)
Transmission	6 (0·54)	44 (0·61)	14 (0·58)	3 (0·33)	67 (0·56)
Place	10 (0·91)	56 (0·78)	19 (0·79)	3 (0·33)	88 (0·76)
Date/time	10 (0·91)	57 (0·78)	17 (0·71)	4 (0·44)	87 (0·75)
Type of online source (n, % origin of online source type)					
Forum/social news/blog	0 (0·0)	5 (0·07)	1 (0·04)	4 (0·44)	10 (0·09)
Reports/press releases	0 (0·0)	6 (0·08)	12 (0·5)	0 (0·0)	18 (0·16)
News	1 (1·0)	61 (0·85)	11 (0·46)	0 (0·0)	83 (0·72)
Cannot classify	0 (0·0)	0 (0·0)	0 (0·0)	5 (0·55)	5 (0·04)

Definitions	
Local/state	Content produced by and for local or state audience
Country	Content produced by and for national audience (may include content by and for local or state regions)
International	Content produced by and for audiences in more than one country
Cannot classify	Unable to determine audience or contains a broad mix
Epidemiologically important information from source	
People	Demographic details and other information that help identify individuals, including names, pronouns, sex, careers, and more
Case history	Information useful for developing model parameters, such as symptom onset and resolution, contact, and more
Transmission	Information about contacts, contact circumstances, and contact events, including names, events, circumstances, and more
Place	Descriptions of locations where individuals traveled, live, work, and more
Date/time	Provides information for when events occurred
Type of online source	
Forum/social news/blog	Content produced by users
Reports & press releases	Content produced by organizations and governments
News	Content produced by a news media organization
Cannot classify	Unable to determine content producer or contains a broad mix

Table 2. *A comparison of Ebola virus disease parameters derived from online sources based largely in Guinea, Liberia, Nigeria, and Sierra Leone as compared with other published estimates*

Source	Mean incubation period (days)	Mean serial interval (days)	Reference
Our estimate	12·5 (95% CI 10·6–14·5)	19·4 (95% CI 17·6–21·3)	
Valencia *et al.*		12	[22]
Van Kerkhove *et al.*	9–12 (mean range)	14–15 (mean range)	[23]
WHO 2014	11·4 (observed); 9·7 (fitted)	15·3	[5]
WHO 2015	10·3	14·2	[24]
Faye *et al.*	9·9	14·2	[25]

Note: The estimates in Van Kerkhove *et al.* are based on a review of other published estimates, including those listed here as WHO 2014, WHO 2015, and Faye *et al.*

Table 3. *Advantages and challenges of using publicly available online resources to build transmission chains*

Advantages	Comments
Timeliness	Data are published often and can be near real-time
Accessible	An internet connection is usually the only requirement
Potentially high granularity	Data can include detailed contact events, onset times of specific symptoms
Understand extrinsic factors	Appreciation for important social, economic, and political factors
Multiple reports of one event	Many viewpoints can enrich and clarify events, and uncover others
Capture of probable events	Can include likely infections not meeting a strict case definition
Reveal otherwise hidden details	Journalists' questions may reveal keys to transmission that might have gone undetected
Challenges	**Comments**
Subject to news competition	Only the most newsworthy and sensational events make the news.
Requires considerable effort	Extract, match, and verify information; requires understanding of indigenous references
Disrespect for privacy	News reports can be intrusive and reveal private details
Loose case definition	Reporters may not be aware of or follow evolving, strict case definitions
Subject to media gags or censorship	Some governments can place a media gag or restrict journalists
Subject to undocumented edits	Online accounts can be altered and removed without warning
Unknown accuracy and precision	Reports can include speculation, misinformation, contradictory information, and even lies.
Capture unpopular details	May include politically/culturally unpopular events that might not be found in official reports

21-day quarantine policies were suitable for the West African outbreak.

DISCUSSION

We have demonstrated that reviewing news reports and online sources originating primarily within the countries ravaged by Ebola in West Africa yielded details sufficient to create transmission trees of infection. In addition, we were able to derive certain disease transmission model parameters that compared favorably with results published elsewhere. Having partial dates, names, or locations, we estimate that approximately two-thirds of the webpages and unique web domains were found to be useful for our purposes of extracting relevant details about the cases of Ebola (Table 1). We conclude that these online details can provide a reasonable means for epidemiologists and disease modelers to derive reliable transmission parameters. In some cases, data derived by this manner may suffice until superior or additional data becomes available through those who are directly involved in the outbreak response. We list advantages as well as challenges of using this method for building a transmission chain (Table 3).

Our approach to gathering epidemiological details via indigenous online sources adds a 'local' news angle to the growing collection of papers reporting similar methods that tap international and specialty online sources. A paper on the Ebola virus disease epidemic by Cleaton *et al.* [13] sourced reports from

https://doi.org/10.1017/S0950268817000760 Published online by Cambridge University Press

US-based news agencies (*The New York Times* and *The Washington Post*), from official WHO publications, and from an online news organization that launched in October 2014 specifically to cover the Ebola virus disease epidemic (*EbolaDeeply,* which hired some reporters with indigenous knowledge of the affected countries). Similarly, a paper by Majumder *et al.* [14] involved assembly of an extensive list of MERS-CoV cases from a 2014 outbreak in Saudi Arabia. While the Majumder paper sourced some smaller stories, it relied heavily on WHO press releases, reports from the Saudi Arabia Ministry of Health, and a paper from the *New England Journal of Medicine*. Furthermore, a supplementary paper by Chowell *et al.* [15], described as an 'exten[sion] and update' of the Cleaton paper, reported successful builds of 104 clusters of Ebola virus disease using WHO reports, situational reports, and 'online authoritative media outlets'. By contrast, we focused our online searches on accounts written by indigenous West African online sources for these same communities as the intended audiences.

Both approaches appear to yield details useful for generating transmission trees and estimating modeling parameters, and each of these approaches has its advantages. The sources used in the Cleaton, Majumder, and Chowell papers were trusted, 'authoritative' sources, including reports generated by the public health organizations themselves (such as the WHO reports), and stories authored by professional journalists who had experience reporting in medical and public health news beats. A potential risk of these sources is that some (WHO situation reports aside) probably existed *because* the Ebola and MERS-CoV epidemics had caught sufficient worldwide attention. It is likely that these international news services and journal-based resources would not have been publishing news during and prior to the critical initial weeks of an outbreak, or for an outbreak that had failed to gather a worldwide audience. An advantage of our approach is that the majority of online sources we used were already publishing stories at the local, state, or national level before the Ebola virus disease epidemic began, persisted through it, and most remain online to this day. While incorporating Ebola virus disease stories into the regular news cycle was new, familiarity with the culture and population would have required few (if any) additional steps.

The transmission chain we constructed for the Nigeria cluster compared favorably with the one reported in the MMWR for the same cluster [12] and similar (though not identical) to the account published by Folarin *et al.*, which was based on genomic and official contact tracing records [16] (we are unaware of any published versions of the cluster at St Joseph's Catholic Hospital in Monrovia, Liberia). Compared with the MMWR report, our first-generation case disease onset times tended to be nearly 1 week later, and were approximately the same for second- and third-generation cases. Our transmission tree differs from that in Figure 2.B reported in Folarin *et al.* in the number of cases (we have 20, they have 19) and in the number of second-generation branches stemming from the index case (we have three, they have two). Both transmission trees show 13 cases linked to the index case, and the branches related to Port Harcourt secondary and tertiary cases are in agreement. We were not able to articulate why we had one more second-generation branch than that stemming from the contact tracing record as reported in Folarin. Of these three secondary branches, two were fairly well established in the online accounts: a professional contact to the Nigeria index case fled to Port Harcourt, and a nurse to the index case had extended contact with her fiancé. The third second-generation branch was harder to establish, though we believe there was sufficient evidence to support the case that a physician who contacted the index case likely also had contact with his wife who later became symptomatic. Explanations for the discrepancies between our account and those recounted in MMWR and Folarin include incorrect recall or recall bias, errors (for example, one story detailed an extremely unlikely account of patient–physician contact, transmission, subsequent symptom onset, and isolation all taking place on 28 July 2014), dissimilar use of the concept of elapsed days, or just incorrect first- or second-hand information. Whereas the contact tracing information from the Folarin paper's supplementary material had precise dates for hospitalization and symptom onset, we sometimes had to infer symptom onset based on mentions of seeking medical attention or hospitalization. Survivors gave the most detailed reports, but they generally needed to be discharged from isolation and spend some time recovering before they could tell their stories to the media. There was an inevitable loss of precision since as much as 2–4 weeks would elapse before survivors could meet with journalists. In practice, this would cause in an important lag or delay before these more precise survivor-based details became available.

Incubation time data from Ebola patients collected from online sources during the outbreak produced estimates of both the mean and 95th percentile that

were remarkably similar to those produced by WHO 2014 study [5], which required detailed investigation of thousands of patients by workers in the affected countries. While internet-derived data cannot replace such investigations, their consistency in this case is promising evidence that internet sources can be used to derive at least reasonably accurate estimates of useful incubation statistics. For example, the estimate of 20·4 days for the 95th percentile incubation time could potentially have provided earlier, concrete reassurance that the >21 day quarantine policy for potentially exposed cases was likely to be reasonable.

Our estimate for the mean serial interval derived from internet sources was higher than the WHO 2014 estimate. A possible explanation for this difference is that the news stories preferentially reported on patients within more explosive portions of the outbreak, including patients who were not identified and thus continued to pose risk of transmission to their contacts well after their symptoms began. Transmissions from these patients would then produce longer serial intervals than transmissions from patients identified earlier. This phenomenon has been observed in data from outbreaks of other diseases: serial intervals early in the 2003 SARS outbreak in Singapore were observed to be longer than those occurring later in the outbreak after control measures were implemented [17]. Similarly, serial intervals during the recent explosive outbreak of MERS in South Korea were observed to be longer than those occurring during sporadic, smaller transmission clusters in the Middle East [18].

Mean serial interval estimates during an outbreak are useful in conjunction with incidence time series for calculating estimates of the effective reproductive number R [5]. Using these methods, an overestimate of the mean serial interval would lead to an overestimate of R [4]. We are unable to determine whether using the methods we describe will regularly result in conservatively high estimates of the reproductive number, or whether that may be a result of the particular cases and clusters we investigated. Future research on this issue may help provide clarification.

We can use our familiarity with the experiences to explain some of the serial interval overestimation. The bulk of the transmissions at St Joseph's Catholic Hospital happened after the hospital's medical director fell ill but was wrongly declared Ebola negative (either a false-negative result or testing was too early). About 5 days before this test, the medical director had direct contact with a symptomatic Ebola patient. Eleven days after the initial negative test, the medical director's health had not improved and he was retested – this time the result was positive for Ebola virus. Unfortunately, following the initial test result (the false negative), many of those caring for the medical director believed he was suffering with malaria and relaxed their use of personal protective equipment. (We included the false-negative and true-positive test events in Fig. 1 for the hospital director.) This event produced relatively high-risk exposures among the hospital staff some 5 days later than what they would have otherwise encountered. Our serial interval estimates were about 4 days longer than those published by the WHO, some of which may be the result of the unique circumstances surrounding the exposures just outlined.

There are inherent strengths to using these types of resources for generating transmission parameters. As demonstrated in the paragraph above, familiarity with event circumstances gleaned through reading transmission accounts may help infectious disease modelers think through anomalies in their data. With respect to locally generated news and online sources, reporters are often among a community's most connected and informed individuals, and the rapid, competitive news cycle means details will often be regular and timely. Reading personal accounts of morbidity and mortality may lead to a deep appreciation and compassion for the people and their circumstances at the community level, including reactions to and perceptions of public health emergencies. Finally, public health agencies (local, national, and more so for international) adopting this approach may gain an understanding of unique cultural and socio-political issues associated with an infectious disease outbreak (e.g. West African funeral practices). This appreciation is not inconsequential: 'It is the population experience of disease, in actual societies, that is the subject of our investigations', write Nancy Kreiger and Sally Zierler in their paper [19] on the nature of explaining the public's health. 'Epidemiologic theory reminds us that our work has a context, and that this context is human society'.

We acknowledge several limitations. The process is resource-intensive and requires adjudicating conflicting details. Sometimes informants speculate, descriptions of people and events can be ambiguous, stories may be updated without highlighting that corrections were made, and newspapers may print inaccurate reports. Reporters may not be aware of or follow official case definitions – especially when those definitions are in flux. It is not always possible to know a reporter's

https://doi.org/10.1017/S0950268817000760 Published online by Cambridge University Press

accidental or deliberate political, cultural, or social biases, or how thoroughly reporters vetted and corroborated their facts. The nature of interest-in-print and privacy means certain important transmission activities (such as sexual transmission of Ebola virus) likely will not be printed. Government censoring or altering stories poses a risk to the timely availability of relevant information. Nearly all of the stories we read were written in English; reports in other languages would likely add another level of complexity that could be addressed by including native speakers on research teams. Not surprisingly, the most sensational and extreme events were the most likely to be reported (journalists sometimes refer to this as the *Man Bites Dog* effect) and also the most detailed, and therefore the easiest to document. The events related to the Ebola cluster in Nigeria, for example, were easy to reconstruct because they generated intense public interest and press coverage: vivid headlines and stories relayed tales of nurses and physicians who died preventing Nigeria's Ebola index patient – who Nigeria's (then) President Goodluck Jonathan called 'a crazy man' – from escaping into Lagos, Nigeria, a city of 21 million people, with one of the world's deadliest infectious diseases. Among the dead were a soon-to-be-married (and pregnant) nurse and a physician who came from one of Nigeria's most prominent families. In this regard, it would be important to perform these searches with respect and compassion. Online newspaper reports appear to be excluded from recent discussions on the ethics of mining social media and big data for public health purposes [20, 21]. Finally, the true extent of the outbreak would likely not be known from newspaper reports.

It is important to consider the generalizability and future applicability of this approach. It is likely that we could reproduce this strategy for another cluster of cases in the recent Ebola outbreak or another public health crisis; the key factor is the availability of person-level details. It would be challenging to predict the level of details available especially as the next outbreak may be in a different region of the world with different social morays, news reporting practices, and expectations of privacy.

In conclusion, we have shown how mining person-level details from local publicly available online sources can yield chains of transmission useful both for general understanding of transmission and also for deriving parameters useful to infectious disease modelers. While acknowledging the tragic nature of the circumstances and with respect for the privacy of patients and the heroic, selfless acts of providers who have risked their lives to control this outbreak, we provide a protocol and guidance for extracting details that would be of benefit to public health.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found at https://doi.org/10.1017/S0950268817000760

ACKNOWLEDGEMENTS

The authors thank Deborah Hofmann for her assistance with this project and the VA Salt Lake City Health Care System (IDEAS Center 2.0 #150HX001240) for administrative support. They acknowledge funding from MIDAS Synthetic Information Systems for Better Informing Public Health Policymakers (5U 01 GM070694-12, PI: Stephen Eubanks, PhD) and the NIH Rocky Mountain Center of Excellence for Biodefense (U 54 AI-065357 NIAID, PI: John Belisle, PhD).

DECLARATION OF INTEREST

None.

ETHICAL STANDARDS

This study was reviewed by the IRB of the University of Utah School of Medicine and was determined to be exempt from IRB oversight. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

DISCLAIMER

The views expressed in this paper are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.

REFERENCES

1. **Yasmin S.** Ebola infections fewer than predicted by disease models. *Scientific American* 2014; 8 December 2014 (www.scientificamerican.com/article/ebola-infections-fewer-than-predicted-by-disease-models).

2. **Centers for Disease C.** 2014–2016 Ebola Outbreak in West Africa. In: Previous Case Counts, 2016 (www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/previous-case-counts.html).
3. **Haas CN.** On the quarantine period for Ebola virus. *PLoS Currents* 2014; **6**: ecurrents.outbreaks.2ab4b76ba7263ff0f084766e43abbd89.
4. **Wallinga J, Lipsitch M.** How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings Biological Sciences/The Royal Society* 2007; **274**: 599–604.
5. **WHO.** Ebola Response Team. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *The New England Journal of Medicine* 2014; **371**: 1481–1495.
6. **Lloyd-Smith JO, *et al.*** Superspreading and the effect of individual variation on disease emergence. *Nature* 2005; **438**: 355–359.
7. **Reich NG, *et al.*** Estimating incubation period distributions with coarse data. *Statistics in Medicine* 2009; **28**: 2769–2784.
8. **Ikeji L.** Photos: Governor Fashola hosts some Ebola survivors. 09/18/2014. In: Linda Ikeji's Blog (http://lindaikeji.blogspot.com/2014/09/photos-governor-fashola-hosts-some.hgml).
9. **Amagiya F.** How I survived Ebola – Dennis Akagha. *Saturday Vanguard* (Nigeria); 09/01/2014.
10. **Obinna C, Olawale G.** Ebola Virus-free country: 15 days of Ebola survivors horror in isolation centre. *Vanguard* (Nigeria); 10/19/2014.
11. **Osakwe F.** Our close shave with death – Ebola survivors. *National Mirror* (Nigeria); 10/15/2014.
12. **Shuaib F, *et al.*** Ebola virus disease outbreak – Nigeria, July-September 2014. *MMWR Morbidity and Mortality Weekly Report* 2014; **63**: 867–872.
13. **Cleaton JM, *et al.*** Characterizing Ebola Transmission Patterns Based on Internet News Reports. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 2015.
14. **Majumder MS, *et al.*** Estimation of MERS-Coronavirus reproductive number and case fatality rate for the Spring 2014 Saudi Arabia Outbreak: insights from publicly available data. *PLoS Currents* 2014; **6**: ecurrents.outbreaks.98d2f8f3382d84f390736cd5f5fe133c.
15. **Chowell G, Cleaton JC, Viboud C.** Elucidating transmission patterns from internet reports: Ebola and middle east respiratory syndrome as case studies. *Journal of Infectious Diseases* 2016; **214**: S421–S426.
16. **Folarin OA, *et al.*** Ebola virus epidemiology and evolution in Nigeria. *The Journal of Infectious Diseases* 2016; **214**: S102–S109.
17. **Lipsitch M, *et al.*** Transmission dynamics and control of severe acute respiratory syndrome. *Science* 2003; **300**: 1966–1970.
18. **Cowling BJ, *et al.*** Preliminary epidemiological assessment of MERS-CoV outbreak in South Korea, May to June 2015. *Euro Surveillance: Bulletin Europeen sur les maladies Transmissibles = European Communicable Disease Bulletin* 2015; **20**: 21163.
19. **Krieger N, Zierler S.** What explains the public's health? – A call for epidemiologic theory. In: Beauchamp D, Steinbock B, eds. *New Ethics for the Public's Health*. New York, NY: Oxford University Press, 1999: pp. 45–49.
20. **Vayena E, *et al.*** Ethical challenges of big data in public health. *PLoS Computational Biology* 2015; **11**: e1003904.
21. **Vayena E, Mastroianni A, Kahn J.** Ethical issues in health research with novel online sources. *American Journal of Public Health* 2012; **102**: 2225–2230.
22. **Valencia C, *et al.*** Network visualization for outbreak response: mapping the Ebola Virus Disease (EVD) chains of transmission in N'Zerekore, Guinea. *The Journal of Infection* 2017; **74**: 294–301.
23. **Van Kerkhove MD, *et al.*** A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. *Scientific Data* 2015; **2**: 150019.
24. **WHO. Ebola Response Team, *et al.*** West African Ebola epidemic after one year—slowing but not yet under control. *The New England Journal of Medicine* 2015; **372**: 584–587.
25. **Faye O, *et al.*** Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. *The Lancet Infectious Diseases* 2015; **15**: 320–326.