Am. J. Trop. Med. Hyg., 86(1), 2012, pp. 39–45
doi:10.4269/ajtmh.2012.11-0597

Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak

Rumi Chunara,* Jason R. Andrews, and John S. Brownstein

Department of Pediatrics, Harvard Medical School, Boston, Massachusetts; Children's Hospital Informatics Program, Division of Emergency Medicine, Children's Hospital Boston, Massachusetts; Division of Infectious Diseases, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Abstract. During infectious disease outbreaks, data collected through health institutions and official reporting structures may not be available for weeks, hindering early epidemiologic assessment. By contrast, data from informal media are typically available in near real-time and could provide earlier estimates of epidemic dynamics. We assessed correlation of volume of cholera-related HealthMap news media reports, Twitter postings, and government cholera cases reported in the first 100 days of the 2010 Haitian cholera outbreak. Trends in volume of informal sources significantly correlated in time with official case data and was available up to 2 weeks earlier. Estimates of the reproductive number ranged from 1.54 to 6.89 (informal sources) and 1.27 to 3.72 (official sources) during the initial outbreak growth period, and 1.04 to 1.51 (informal) and 1.06 to 1.73 (official) when Hurricane Tomas afflicted Haiti. Informal data can be used complementarily with official data in an outbreak setting to get timely estimates of disease dynamics.

INTRODUCTION

Amidst existing poor healthcare, water, and sewage infrastructure, Haiti suffered a devastating earthquake in January 2010. The combination of these factors left the country vulnerable to the emergence of cholera for the first time in a century. There have been over 380,000 reported cases and 5,800 reported deaths (as of July 10, 2011). Sequencing of *Vibrio cholerae* isolates from the outbreak showed that the epidemic was likely the result of the introduction, through human activity, of a *V. cholerae* strain from a distant geographic source.[1,2] The initial cases were reported in the Artibonite department in Haiti and subsequently spread to all 10 administrative departments. In addition, the same strain of cholera was detected in other countries, including the neighboring Dominican Republic and the United States within 28 days[3,4] with subsequent spread to Venezuela, Mexico, Spain, and Canada. Although control measures for the epidemic have been since initiated, the unique instigation and geographic spread of this epidemic highlight the need for improvements in country-level and global outbreak surveillance for the increasing number and types of infectious disease events around the world.

The Haitian Ministry of Public Health (Ministère de la Santé Publique et de la Population, MSPP) has published data facilitating studies examining the evolution of the epidemic.[5] Groups have used this official data to estimate key epidemic parameters and simulate the impact of preventive and reactive interventions to control disease spread over time.[6–8] Retrospective analyses dependent on data reporting from public health sources are often temporally limited; alternative data sources may provide an opportunity to collect early information about how an epidemic is unfolding, and thus the opportunity for the implementation of more timely and effective interventions.

Here, we investigate the use of alternative data sources for understanding disease epidemiology. The Internet has become one of these sources, used ubiquitously by a variety of groups including clinicians, public health practitioners, and laypeople, to seek health information. In addition, the Internet serves as an accessible reservoir for the public regarding official announcements disseminated by government agencies and informal news from press reports, blogs, chat rooms, web searches, and media reports.[9] In particular, volume of some Internet metrics such as web searches or microblogs have been shown to be a good corollary for public health events.[10,11] In this study, we evaluate trends in the volume of online social and news media to determine whether they correlate with officially reported disease measures, and we show their potential use in estimating a key epidemic parameter.

METHODS

Data sources. We examined data from the first 100 days (October 20, 2010 through January 28, 2011) of the Haitian cholera outbreak from three sources: HealthMap, Twitter, and MSPP. Here, we refer to data from HealthMap and Twitter, which are unvetted by government or multilateral bodies such as the World Health Organization (WHO), as "informal." We refer to data from the MSPP, a government body, as "official."

Overview of MSPP data. Since the first weeks of the cholera outbreak in Haiti, the MSPP has published official data[5]; this data includes daily tallies of cholera cases, cholera-related hospitalizations, and deaths, reported by department. As is common with information requiring officiating through multiple organizational levels and in a complex situation such as a disaster, the MSPP daily data regarding the Haiti cholera outbreak are published in batches, and can appear with a time lag anywhere from ~7 to 14 days. Reports are updated retrospectively with new reports updating counts from previous days as the information being sourced changes. For this analysis, we used the total cholera cases seen as reported by the MSPP (Supplemental Figure S1).

Overview of HealthMap and Twitter data. The first informal source we examined was news media volume acquired via HealthMap (see: http://www.healthmap.org). HealthMap is an automated surveillance platform that continually identifies, characterizes, and maps events of public health and medical importance, including outbreaks and epidemics.[12] Information sources for HealthMap include news media sources and

*Address correspondence to Rumi Chunara, 1 Autumn St., Suite 433, Boston, MA 02215. E-mail: rumi@alum.mit.edu

discussion groups. HealthMap also incorporates data from the community through the "Outbreaks Near Me" mobile phone application,[13] wherein any user may contribute reports via their phone, and online contributions on its website. From October 20 and continuing until approximately November 17, an extended effort regarding the cholera outbreak in Haiti was conducted to find and supplement information already gathered through HealthMap. This involved an increase in active surveillance, especially in French language feeds, and work with partners in the United States and on the ground in Haiti (Humanity Road, Ushahidi, Harvard Humanitarian Initiative, Crisis Mappers, School of International and Public Affairs at Columbia, OpenStreetMap team, International Organization for Migration, MissionMANNA). HealthMap is routinely updated automatically, and additional data garnered about the outbreak were added into the HealthMap system by human curators using an administrator tool.

During the epidemic we created a map that was updated in real-time (http://www.healthmap.org/haiti), with additional layers of pertinent relief information such as the locations of hospitals, cholera treatment centers, new safe water installations, and water points. Figure 1 illustrates the time course of the Haiti HealthMap alerts obtained from October 20 to January 28 (100 days). As shown in Figure 1, it is possible to view the HealthMap data grouped into a level such as the 41 arrondissements or at the level of precise location, which can allow for finer understanding of where disease activity is occurring during an outbreak.

The second informal source examined was cholera-related postings on the website Twitter (http://www.twitter.com). Twitter is a microblogging service in which users can give information in 140 character length posts, referred to as "Tweets." We collected historical Tweets for the chosen date range via the Research.ly (http://www.research.ly) interface on March 4, 2011. We selected all publically available Tweets containing the word "cholera" including those with the Twitter hashtag identifier ("#cholera"). Our search captured English and French mentions of the word "cholera" as well as Tweets in all languages. Tweets regarding cholera existed before the start of the cholera outbreak in Haiti, but commencing with the start of the outbreak, Tweets containing the word cholera that did not have to do with Haiti were sparse, thus we used the global volume of Tweets. Content of the Tweets included personal concerns from family or friends, local happenings on the ground, and reiteration of news reports. "Sitting w/a father who just lost his 7-year old to cholera. Reality still has not hit," "My visit to Saint Nicolas hospital in Saint Marc, As Haiti is still fighting Cholera" are example messages posted on Twitter early in the outbreak.

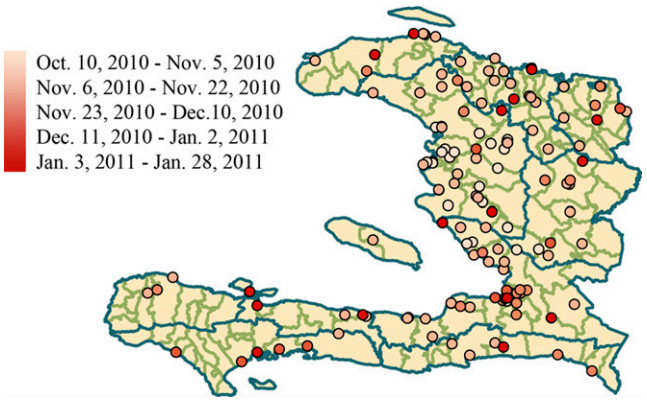

FIGURE 1. Time and space distribution of HealthMap alerts for the first one hundred days of the Haiti cholera outbreak. Each HealthMap alert (marked by a circle colored corresponding to its date) is precise placed to an exact latitude and longitude, and could also be generalized to the administrative areas it falls within. The 10 departments (largest administrative jurisdiction) of Haiti and 41 arrondissements (next largest administrative jurisdiction) are outlined in dark and light borders, respectively.

Analytical framework. To evaluate the use of informal sources for understanding the epidemic over time, we first examined three major time periods of activity for correlation between the curves, which are illustrated in Figure 2. The first time period is during the initial phase of the outbreak, October 20 (the date of the first officially reported case data from the MSPP) to November 3 (the date when the original peak of cases subsides). The second time period represents the increase in cases around the timing of Hurricane Tomas. Tomas was first classified as a tropical storm on October 29, and was at Hurricane status when it passed closest to Haiti on November 5.[14] We chose November 3–December 1 to encompass the anticipation of this event, the event, and repercussions, which also corresponded to the second peak of cases. Finally, we chose the 100 days from the start of the outbreak as the third time period. Of note, Twitter data was unavailable from January 25 to January 30, and therefore contained six fewer data points.

To understand the temporal relationship between the HealthMap, Twitter, and MSPP data, we computed the Pearson's cross-correlation coefficient, ρ, between each pair of the unfiltered data sources. Examination of autocorrelation and partial autocorrelation function plots of the raw data did not dictate any prefiltering necessary to account for underlying trends. This metric measures the strength of the linear association of two waveforms as a function of a time lag applied to

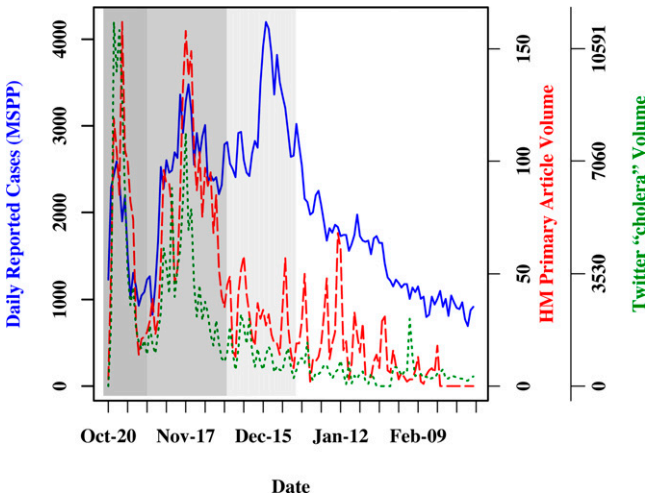

FIGURE 2. Daily reported case data for all departments from the Haiti Ministry of Health (solid), daily volume of primary HealthMap alerts (dashed), and daily volume of Twitter posts containing the word "cholera" or "#cholera" (dotted). Each curve has an initial peak at the onset of the outbreak (dark grey), and a peak during the time that Hurricane Tomas affected Haiti (medium grey). The first 100 days of the outbreak are shaded in light grey. Ministère de la Santé Publique et de la Population (MSPP) case counts peak again in late December, although HealthMap and Twitter volume only have daily variations during this time.

one of them, and does not depend on the units in which the two variables are measured[15]; temporal resolution was at the level of days.

Next, we created epidemic curves from the cumulative volume of informal data (from each of HealthMap and Twitter) and MSPP-reported case burden to make estimates of the effective reproductive number from each source. The effective reproductive number, R_e, is the mean number of secondary cases generated by the average infectious individual after an epidemic has begun. This important epidemic parameter can be estimated using the Lotka-Euler equation,[16] which relates R_e to the distribution of mean serial interval (the time between infections in consecutive generations) and the epidemic growth rate. Early in an epidemic, the effective reproductive number closely approximates the basic reproductive number. We estimated the growth rate from the epidemic curves during two phases of exponential growth that occurred in all three data sources: an initial period of rapid growth as the cholera epidemic originally spread (phase 1, ~October 20–30), and another period of rapid disease spread during flooding experienced after Hurricane Tomas (phase 2, ~November 7–19) (Figure 3).

Estimates for the serial interval of cholera can vary highly because of environmental factors influencing the disease's spread. Cholera can be transmitted from person to person through contaminated food or household water sources, or through environmental aquatic reservoirs, where it may survive for weeks to months. Person-to-person spread would be associated with a serial interval on the order of a couple days, whereas environmental transmission may be associated with serial intervals of weeks. Studies of household transmission do not necessarily indicate person-to-person transmission, but nevertheless provide the best data on time between observed cases. We used a range of 1–9 days, which was supported by clinical evidence from household studies of cholera transmission[17–20] (Supplemental Table S1). In addition, for understanding the distribution of the mean serial interval, we combined data from two of the household studies reporting time between appearance of initial and secondary cases in a household, and fit the curve of days between cases to an exponential distribution (exponential fit, $R^2 = 0.800, 0.635$, respectively).[18,21] We also addressed the possibility of transmission through purely environmental routes, with longer serial intervals, by extending our sensitivity analysis to mean serial intervals up to 30 days (Supplemental Figure S2).

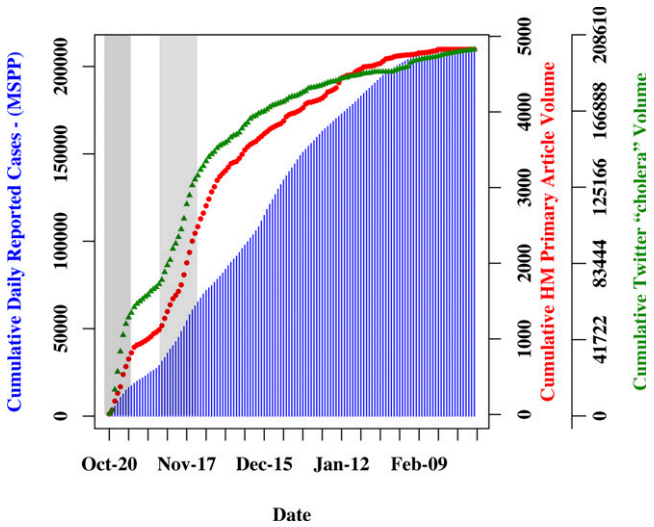

FIGURE 3. Epidemic curve generated from cumulative histogram of MSPP case counts from all departments (bars), cumulative number of primary HM articles (circles), and cumulative number of cholera-Twitter posts (triangles). Dark grey highlights the first period of rapid growth, light grey the second.

For an exponentially distributed mean serial interval time with mean $T_c = 1/b$ (where b is the rate of leaving the infectious stage in a susceptible-infectious-recovered model), there is a linear relationship between growth rate, r, and the reproductive number[16]:

$$R_e = 1 + rT_c. \quad (1)$$

The latency period of cholera is small (on the order of hours[22]) compared with the serial interval, allowing us to ignore latency in these analyses (Supplemental Figure S3).

Growth rates were measured from periods of exponential growth in each epidemic curve, allowing for the longest amount of time in which exponential growth occurred in each phase (10 and 12 days for phases 1 and 2, respectively). The reproductive number was calculated from each data source using the measured growth rate and the selected range of serial intervals. Error for our estimates of R_e were calculated through propagation of uncertainty from the growth rate and serial interval parameters. Error from the growth rate was determined by the 95% confidence interval (CI) from the exponentially fitted parameter. For the serial interval parameter, error was determined based on the standard deviation of the selected distribution. Statistical analyses were conducted using the statistical software R, version 2.130 (R Foundation, Vienna, Austria).

RESULTS

Comparison of MSPP and informal data. Cases of cholera in Haiti were first confirmed by October 19.[1] The first case and hospitalization reports from the MSPP are from October 20. On the same day there were three HealthMap alerts reporting deaths in the preceding days from a diarrhea outbreak and suspected cholera. News articles confirming cholera appeared on October 21; on this day there were 51 HealthMap alerts, followed by hundreds in each of the subsequent days. The first cholera Tweets were on October 21; on this day there were 1,995 HealthMap alerts, and thousands on each of the subsequent days.

Initially, we compared total HealthMap volume over time to MSPP data. There was a large spike at the outset of the cholera outbreak, and during the initial days of influence of Hurricane Tomas, which was anticipated to and did increase cholera cases in Haiti and surrounding regions. To account for this media spike we filtered the histogram to incorporate only "primary" articles (an article in the HealthMap database is deemed primary if it is the first article containing new information, and subsequent articles containing the same information are "children" of that article). The total article volume peaks occurred concurrently with peaks of the primary article volume during this study. Figures 3 and 4 incorporate this filtered, "primary" HealthMap article volume over time.

HealthMap and Twitter data both showed distinct peaks at the outset of the outbreak (from October 20 to November 3 there were 995 primary HealthMap alerts about the outbreak and

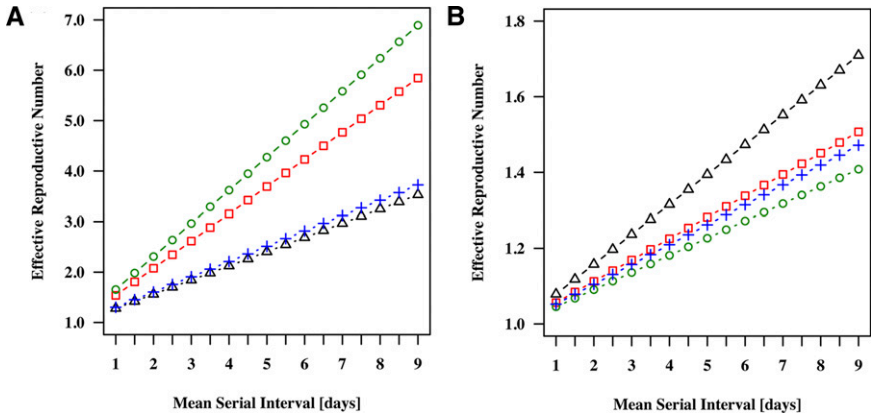

FIGURE 4. Estimates of the effective reproductive number. Estimates are for a range of plausible serial intervals assuming 100% person-to-person transmission, an exponential distribution of the mean serial intervals and negligible latency period, for each of the data sources (crosses: Ministère de la Santé Publique et de la Population [MSPP] cases, triangles: MSPP hospitalizations, squares: HealthMap primary alerts, circles: "cholera" Tweets). Growth rate was extracted through an exponential fit to each data source for both phases (**A**, phase 1, October 20–30; **B**, phase 2, November 7–19).

65,728 cholera-related Tweets), and during the period of influence by Hurricane Tomas (from November 3 to December 1, 2,248 HealthMap alerts, 84,992 Tweets), which correspond to the trend observed in the MSPP case volume.

In total for the time period between October 21 and January 28, we captured 188,819 tweets and 4,697 HealthMap primary alerts. There was also a third peak in the MSPP data around the second week of December that did not appear in the informal data (Figure 2).

Temporal correlation of data. Informal data sources (HealthMap, Twitter) had the highest correlation in time at 0 days lag for all three time periods (October 20–November 3: ρ = 0.81, 95% CI = 0.64–0.91; November 3–December 1: ρ = 0.77, 95% CI = 0.64–0.86; and October 20–January 28: ρ = 0.80, 95% CI = 0.75–0.84) (Table 1). In comparing MSPP hospitalization data with HealthMap data, a better correlation was observed in the first two time periods (October 20–November 3: ρ = 0.76, 95% CI = 0.55–0.88; November 3–December 1: ρ = 0.76, 95% CI = 0.63–0.85), with poorer correlation in the third time period (October 20–January 28: ρ = 0.41; 95% CI = 0.29–0.51). Similarly, MSPP hospitalization data had higher correlation with Twitter data in the first two periods (October 20–November 3: ρ = 0.86, 95% CI = 0.71–0.93; November 3–December 1: 0.57, 95% CI = 0.36–0.72), but lower in the third period (October 20–January 28: ρ = 0.25; 95% CI = 0.13–0.37). The MSPP case data precedes the HealthMap and Twitter day by 1 day (best correlation with data lagged by 1 day).

Effective reproductive number estimates. Figure 4 shows how estimates of R_e vary based on the chosen mean serial interval. For our estimated range of the mean serial interval, estimates of R_e using the informal sources ranged from 1.54 to 6.89 (Figure 4A) and 1.04 to 1.51 (Figure 4B), and using the official data were between 1.27 and 3.72 (Figure 4A) and 1.06 and 1.73 (Figure 4B). Thus, the R_e estimates were most similar in phase 2 (estimates from official and informal sources differed by 1.9–14.6%), and for smaller mean serial interval assumptions in phase 1 (20% at mean serial interval of 1 day). We also examined how these estimates vary based on the number of days used to extract the growth rate during each phase (Figure 5). Earliest estimates of R_e from informal data sources would be up to 3.5 times larger than those made from observation of data through the end of the exponential growth period. Estimates of R_e from official data were up to 2.4 times higher when using data from the initial portion of the epidemic compared with data through the entire exponential growth period. Error of the R_e estimates are illustrated in Supplemental Figure S4.

DISCUSSION

In the early days of a disease outbreak, clinicians, public health officials, and policy makers need rapidly available data to plan a response to an impending epidemic. Data collected and reported through official public health institutions is often not available for weeks while reporting mechanisms are established

TABLE 1
Cross-correlations between time series of the three data sources

Data source 1	Data source 2	Date range	Correlation 0 days lag (95% CI)	Correlation 1 day lag (95% CI)
MSPP cases	HealthMap	October 20–November 3	0.66 (0.39–0.87)	0.76 (0.55–0.88)
		November 3–December 1	0.71 (0.55–0.82)	0.76 (0.63–0.85)
		October 20–January 28	0.39 (0.28–0.49)	0.41 (0.29–0.51)
MSPP cases	Twitter	October 20–November 3	0.83 (0.66–0.91)	0.86 (0.71–0.93)
		November 3–December 1	0.57 (0.35–0.72)	0.57 (0.36–0.72)
		October 20–January 28	0.25 (0.13–0.37)	0.25 (0.13–0.37)
Twitter	HealthMap	October 20–November 3	0.81 (0.64–0.91)	0.67 (0.40–0.83)
		November 3–December 1	0.77 (0.64–0.86)	0.62 (0.42–0.76)
		October 20–January 28	0.80 (0.75–0.84)	0.75 (0.68–0.80)

CI = confidence interval; MSPP = Ministère de la Santé Publique et de la Population.

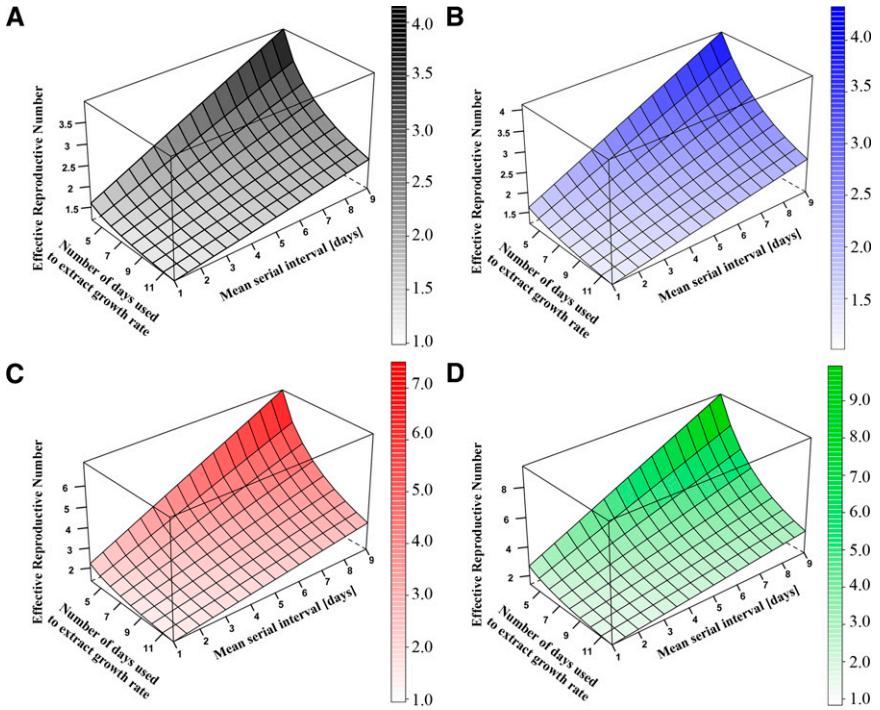

FIGURE 5. Variation of effective reproductive number based on the number of days of data used to calculate the growth rate. Change in estimates of the effective reproductive number in phase 1 (October 20–30) as the number of days used to extract growth rate during exponential growth periods is varied (**A**, HealthMap; **B**, Twitter; **C**, Ministère de la Santé Publique et de la Population [MSPP] cases; **D**, MSPP hospitalizations). Growth rate was extracted from the epidemic curves using from 4 to 12 days of data and the serial interval was varied from 1 to 9 days as in Figure 4.

and bolstered. We examined data from two informal sources—HealthMap and Twitter, made available on the Internet in real-time, to determine whether the trend in volume over time of such reports would correlate with the trend in volume of cases reported through official mechanisms over time. We found that in the 2010 Haitian cholera outbreak, there was good correlation between trends in volume over time of informal data and officially reported case data, during the initial stages of an outbreak or relevant event. We demonstrate one potential use of this informal data early in an outbreak to gain early insight into an evolving epidemic—estimating the reproductive number of the cholera epidemic, which has important implications for the implementation of disease control measures.

Informal media sources such as search query volume have previously been shown to be accurate metrics for "predicting present activity" in economics, sales, disease prevalence, and consumer activity.[10,23] Here for the first time, we show their use in monitoring an outbreak of a neglected tropical disease in a resource-limited setting and in estimating the effective reproductive number of an epidemic, to gain early insight into disease dynamics.

We found that data from the informal sources correlated best with MSPP data with a 1 day lag, meaning that the changes in volume occurred 1 day later in those sources than in MSPP data. However, these data sources are made publicly available in real-time, whereas MSPP data is released with up to 2 weeks of delay. Thus, because access to informal sources is possible in near real-time, estimates from these data sources can be made earlier than from formal sources, which are available after delays incurred in the traditional chain-of-command structure of public health. The use of electronic sources can also facilitate finer temporal resolution than more traditional data streams; often at the level of single days or better. Consequently, estimates derived from these data sources can be generated very early and often, with the potential to precede insight available from official sources. Electronic sources also offer very fine spatial resolution, which is not explored in this study. Near real-time estimates of epidemic activity may provide valuable insights into the trajectory of an infectious disease outbreak, help project the spread of an epidemic, and provide guidance on the magnitude of control measures needed. The reproductive number can be used to determine the proportion of the population that needs to be immunized to contain an epidemic, or the proportion that will be infected when the disease reaches its endemic equilibrium.

In the study presented here, we found that trends in the volume of informal media sources correlated with trends in official case volume early in the epidemic, during periods of exponential growth, where estimates of R_e are made. We showed how estimation of R_e can vary based on the number of days used to determine this growth rate. Very early estimates from media sources diverged much more than early estimates from official data, indicating a media amplification effect around initial news of an event. During the second period of exponential growth (around the time of Hurricane Tomas), the growth rates were very similar for informal sources and official sources. This could suggest that the media amplification effect may be more important around the time of a new outbreak, whereas this phenomenon is less relevant as an epidemic continues to spread. Because epidemic curves from informal sources had exponential growth during corresponding time periods, estimates of the reproductive number could be made within 10 days of the outbreak onset. Although correlation was not good later in the epidemic, it was strong during

periods of exponential rise in cases, which is where the reproductive number is estimated.

In the data from the MSPP, there was a third peak of cases (Figure 2) that was not captured by the informal sources, which could be caused by local disease dynamics that did not garner further media attention, or a loss of media attention after initial stages of the outbreak. Accordingly, the methods here are primarily useful for evaluating the relationship between informal and official data streams during periods of high disease transmission activity, which commonly occurs at the beginning of an outbreak. We found that estimates of R_e using informal sources during the second phase of exponential growth in the Haiti epidemic matched within the calculated error margins for the selected range of mean serial intervals, whereas estimates in the initial phase were larger by ~1.2–1.9× than estimates from official sources. Temporal differences in the relationship between informal and official data streams could be caused by differences in accuracy of official reports or in characteristics of the disease dynamics or media as the epidemic progressed.

In principle, the methods and data types presented here can be extended to other diseases and to other metrics of disease activity. Media sources can act as an independent metric for gauging disease activity, which is unaffected by biases of, or can convey trends not captured in, official data. Passive surveillance data collected from health facilities by the government can be afflicted by temporally varying logistical or political limitations and generally result in underestimation of the true disease burden in epidemics.[2,7,24] Furthermore, for diseases transmitted purely from person to person, the mean serial interval may be better understood allowing for more precise estimates of R_e. For cholera, there is poor data on timing of transmission between individuals, which represents a major source of uncertainty in the estimates of R_e presented here. Alternative approaches for estimating the reproductive number, such as Bayesian or maximum likelihood frameworks, could also be used to estimate the reproductive number early in an epidemic by using informal data volume combined with assumptions about the serial interval distribution.[25–27]

Informal data sources may contain biases that should be considered. First, there may be geographic biases constraining media prevalence; media may be more ubiquitous in and about larger urban centers or developed regions in general. Furthermore, media volume originating from Haiti or any post-disaster environment may be reduced because of poor existing or resulting infrastructure. As well, global media coverage regarding neglected tropical diseases may be reduced even if case burden is similar to events for other diseases. Second, data contributed by individuals from informal mediums (such as microblogging, cell phones, etc.) may be more prevalent from certain age or other demographic groups.[28] However, penetration and use of consumer technology is constantly increasing and facilitating more communication in a variety of worldwide settings, which will decrease demographic and geographic biases in the information. A third potential bias is that informal media reports may contain false positives; they may appear in the absence of disease, based upon false alerts, rumors, or misreporting, particularly in situations of fear or panic. This would contribute to disproportionality between trends in media reports and the underlying volume of disease. Other studies have generated rules for determining relevancy for and reconciling these spikes in time-series data,[29] and these methods could be incorporated into future work. Additionally, broadening our inclusion criteria to include Tweets that also contained words such as "diarrhea" or "vomiting" would have increased the sensitivity of captured Tweets, but decreased the specificity. Finally, we found that correlation between informal media sources and case numbers was not significant later in the epidemic, which may be an important limitation of this method late in epidemics.

We have shown here that social and news media sources yielded data that correlated well with officially reported data from the MSPP. Furthermore, at the early stages of an outbreak informal sources can be indicative not just that an outbreak is occurring, but can highlight disease dynamics through estimation of a key epidemic parameter, the reproductive number. Social and news media such as from HealthMap and Twitter are a cost-effective data source. Further research is needed to determine if informal media will be a good measure of morbidity in other epidemics, and how such sources can best be used for monitoring and characterizing future infectious disease epidemics. The next steps would also entail studying how this could be done prospectively. These methods are not a replacement for traditional surveillance methods; however, our results show that these sources can be used to complement current methods for early estimation of epidemiological parameters.

Received September 23, 2011. Accepted for publication November 8, 2011.

Note: Supplemental figures and tables appear at www.ajtmh.org.

Acknowledgments: We acknowledge Clark Freifeld, Amy Hansen, and Sumiko Mekaru for help accumulating the HealthMap alerts used here, and Emily Chan for advice on data analysis.

Financial support: Financial support for this study was provided by research grants from Google.org, the National Library of Medicine (5G08LM9776-2), and National Institutes of Health (1R01LM01812-01) to RC and JSB, and National Institute of Allergy and Infectious Diseases training grant (T32AI007433-20) to JRA.

Authors' addresses: Rumi Chunara and John Brownstein, Children's Hospital Informatics Program, Harvard Medical School, Boston, MA, E-mails: rumi@alum.mit.edu and john.brownstein@childrens.harvard.edu. Jason Andrews, Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, E-mail: jandrews6@partners.org.

REFERENCES

1. Chin C, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK, 2011. The origin of the Haitian cholera outbreak strain. *N Engl J Med 364:* 33–42.
2. Macdonald G, 1952. The analysis of equilibrium in malaria. *Trop Dis Bull 49:* 813–829.
3. CNN U.S., 2010. Florida woman diagnosed with cholera. Available at: http://www.cnn.com/2010/US/11/17/florida.haiti.cholera/?hpt=T2. Accessed November 17, 2010.
4. BBC News Latin America and Caribbean, 2010. Haiti cholera reaches Dominican Republic. Available at: http://www.bbc.co.uk/news/world-latin-america-11771109. Accessed May 10, 2011.
5. Republic of Haiti Ministry of Population and Public Health, 2011. Documentation on Cholera. Available at: http://www.mspp.gouv.ht/site/index.php?option=com_content&view=article&id=57&Itemid=1. Accessed July 11, 2011.
6. Andrews JR, Basu S, 2011. Transmission dynamics and control of cholera in Haiti: an epidemic model. *Lancet 377:* 1248–1255.

7. Tuite AR, Tien J, Eisenberg M, Earn DJ, Ma J, Fisman DN, 2011. Cholera epidemic in Haiti, 2010: using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Ann Intern Med 154:* 593–601.
8. Chao DL, Halloran ME, Longini IM Jr, 2011. Vaccination strategies for epidemic cholera in Haiti with implications for the developing world. *Proc Natl Acad Sci USA 108:* 7081–7085.
9. Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, Buckeridge DL, 2010. Information technology and global surveillance of cases of 2009 H1N1 influenza. *N Engl J Med 362:* 1731–1735.
10. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L, 2009. Detecting influenza epidemics using search engine query data. *Nature 457:* 1012–1014.
11. Signorini A, Segre A, Polgreen P, 2010. Using Twitter to estimate H1N1 activity. International Society of Disease Surveillance 9th Annual Conference, 1–2 December 2010, Park City, Utah.
12. Brownstein JS, Freifeld CC, Reis BY, Mandl KD, 2008. Surveillance sans frontières: internet-based emerging infectious disease intelligence and the HealthMap Project. *PLoS Med 5:* e151
13. Freifeld CC, Chunara R, Mekaru SR, Chan EH, Kass-Hout T, Ayala Iacucci A, Brownstein JS, 2010. Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS Med 7*: e1000376.
14. National Hurricane Center, 2010. Hurricane Tomas Discussion Twenty-Nine. Available at: http://www.nhc.noaa.gov/archive/2010/al21/al212010.discus.029.shtml? Accessed May 10, 2011.
15. Chatfield C, 2004. *The Analysis of Time Series: An Introduction*. New York: Chapman and Hall/CRC.
16. Wallinga J, Lipsitch M, 2007. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc Biol Sci 274:* 599–604.
17. Kendall EA, Chowdhury F, Begum Y, Khan AI, Li S, Thierer JH, Bailey J, Kreisel K, Tacket CO, LaRocque RC, Harris JB, Ryan ET, Qadri F, Calderwood SB, Stine OC, 2010. Relatedness of *Vibrio cholerae* O1/O139 isolates from patients and their household contacts, determined by multilocus variable-number tandem-repeat analysis. *J Bacteriol 192:* 4367–4376.
18. Mosley WH, Ahmad S, Benenson AS, Ahmed A, 1968. The relationship of vibriocidal antibody titre to susceptibility to cholera in family contacts of cholera patients. *Bull World Health Organ 38:* 777–785.
19. Weil AA, Khan AI, Chowdhury F, Larocque RC, Faruque AS, Ryan ET, Calderwood SB, Qadri F, Harris JB, 2009. Clinical outcomes in household contacts of patients with cholera in Bangladesh. *Clin Infect Dis 49:* 1473–1479.
20. Rahman KM, Duggal P, Harris JB, Saha SK, Streatfield PK, Ryan ET, Calderwood SB, Qadri F, Yunus M, LaRocque RC, 2009. Familial aggregation of *Vibrio cholerae*-associated infection in Matlab, Bangladesh. *J Health Popul Nutr 27:* 733–738.
21. Cavanaugh DC, Thorpe BD, Bushman JB, Nicholes PS, Rust JH Jr, 1965. Cholera in East Pakistan families, 1962–63. Pakistan-Seato Cholera Research Laboratory. *Bull World Health Organ 32:* 205–209.
22. Killeen TS, Waldor KP, Beattie MK, Spriggs DT, Kenner DR Jr, Trofa A, Sadoff JC, Mekalanos JJ, Taylor DN, 1995. Safety, immunogenicity, and efficacy of live attenuated *Vibrio cholerae* O139 vaccine prototype. *Lancet 345:* 949–952.
23. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ, 2010. Predicting consumer behavior with Web search. *Proc Natl Acad Sci USA 107:* 17486–17490.
24. Maskalyk J, Hoey J, 2003. SARS update. *CMAJ 168*: 1294–1295.
25. Cauchemez S, Boelle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, Hedley AJ, Anderson RM, Valleron AJ, 2006. Real-time estimates in early detection of SARS. *Emerg Infect Dis 12:* 110–113.
26. Wallinga J, Teunis P, 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol 160:* 509–516.
27. White L, Wallinga J, Finelli L, Reed C, Riley S, Lipsitch M, Pagano M, 2009. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza Other Respi Viruses 3:* 267–276.
28. Nielsen, 2010. *Mobile Youth Around the World*. New York: The Nielsen Company.
29. Chan EH, Sahai V, Conrad C, Brownstein JS, 2011. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis 5:* e1206.