

Predicting social response to infectious disease outbreaks from internet-based news streams

Shannon M. Fast¹ · Louis Kim¹ · Emily L. Cohn² · Sumiko R. Mekaru² · John S. Brownstein² · Natasha Markuzon¹

Published online: 29 March 2017
© Springer Science+Business Media New York 2017

Abstract Infectious disease outbreaks often have consequences beyond human health, including concern among the population, economic instability, and sometimes violence. A warning system capable of anticipating social disruptions resulting from disease outbreaks is urgently needed to help decision makers prepare appropriately. We designed a system that operates in near real-time to identify and predict social response. Over 150,000 Internet-based news articles related to outbreaks of 16 diseases in 72 countries and territories were provided by HealthMap. These articles were automatically tagged with indicators of the disease activity and population reaction. An anomaly detection algorithm was implemented on the population reaction indicators to identify periods of unusually severe social response. Then a model was developed to predict the probability of these periods of unusually severe social response occurring in the coming week, 2 and 3 weeks. This model exhibited remarkably strong performance for diseases with substantial media coverage. For country-disease pairs with a median of 20 or more articles per year, the onset of social response in the next week was correctly predicted over 60% of the time, and 87% of weeks were correctly predicted. Performance was weaker for diseases with little media coverage, and, for these diseases, the main utility of our system is in identifying social response when it occurs, rather than predicting when it will happen in the future. Overall, the developed near real-time prediction approach is a promising step toward developing predictive models to inform responders of the likely social consequences of disease spread.

This research was funded by Defense Threat Reduction Agency (www.dtra.mil) contact HDTRA1-12-C-0061.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10479-017-2480-9>) contains supplementary material, which is available to authorized users.

✉ Natasha Markuzon
nmarkuzon@draper.com

¹ Information and Decision Systems Division, The Charles Stark Draper Laboratory, Cambridge, MA, USA

² Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Keywords Biosurveillance · Social response · Epidemics · Anomaly detection · Near real-time prediction

1 Introduction

Despite progress in the fight against infectious diseases, they remain a persistent threat to global health, claiming approximately 9.5 million lives annually (Lozano et al. 2012). Moreover, the consequences of disease outbreaks extend beyond human health. Societal strain—ranging from anxiety and economic effects (Cheng 2004) to riots, violence, or flight (Kinsman 2012)—frequently accompanies the outbreak of severe infectious disease. These social responses may ultimately impact national security and can limit responders' ability to combat the disease, as recently observed with the Ebola epidemic in West Africa (International Federation of Red Cross and Red Crescent Societies 2015). A warning system capable of anticipating the social consequences of epidemics will benefit decision makers and relief workers, helping them to allocate resources and respond appropriately. In this work, we present such a warning system and demonstrate its utility for predicting social response to disease outbreaks around the world.

As social media and Internet news data are becoming increasingly prevalent, forecasting of social phenomena using these data has become an area of great interest. Social media and news data streams have been used to predict targets ranging from election results (Gayo-Avello 2013) and financial markets (Bollen et al. 2011; Schumaker and Chen 2009) to urban crime (Gerber 2014) and civil unrest (Montgomery et al. 2012; D'Orazio and Yonamine 2015). Prominent systems, such as the Integrated Crisis Early Warning System (ICEWS) (O'Brien 2010), the Global Database of Events, Location, and Tone (GDELT) (Racette et al. 2014), Early Model Based Event Recognition Based on Surrogates (EMBERS) (Doyle et al. 2014), and Recorded Future (Truvé 2013), harvest data streams from international, regional and local news sources, as well as social media and Internet forums, in order to forecast major political instability events, society-level behavior, and cyber threats. In the field of public health, several systems, including the Global Public Health Intelligence Network (GPHIN) (Mykhalovskiy and Weir 2006), HealthMap (Brownstein et al. 2008), ProMED-mail (Woodall 2001), and Biocaster (Collier et al. 2008), have been developed to facilitate outbreak detection and monitoring. These systems monitor data streams for disease-specific events. Social reactions are frequently discussed in news streams covering disease outbreaks, and predicting the occurrence of social response that might disrupt response efforts is a natural next step for global disease monitoring systems.

Social response to disease outbreaks is a relatively new area of interest for research, with studies primarily focusing on local events. Research has been conducted on the type, timing, and cause of social response for specific disease outbreaks (Sherlaw and Raude 2013; Lau et al. 2010), including the 2003 SARS outbreak in Hong Kong (Cheng 2004) and the 2000–2001 Ebola outbreak in Uganda (Kinsman 2012). Analysis of infectious disease outbreaks with and without social response has revealed that severe social response occurs most frequently when pathogens are clinically severe or are novel to local experts (Fast et al. 2015; McGrath 1991), and that countries with low per-capita health expenditure and high levels of armed conflict and child mortality may be particularly susceptible (Vaisman et al. 2014).

In the current work, we extend these efforts, laying the groundwork for a near real-time warning system for social response. The method provides forecasts of the social response

for the coming 1, 2, and 3 weeks. The model's primary data source was a collection of Internet-based news articles from the HealthMap historical database and daily data stream. HealthMap (Brownstein et al. 2008), in operation since 2006, aggregates epidemic intelligence from multiple data sources, including news, social media, crowdsourced intelligence, and formal reports to identify health events, often prior to formal investigations. It has been shown that information derived from Internet-based news sources provides early and accurate information for disease detection and analysis of spread (Wilson and Brownstein 2009), but the utility of such information for predicting social response has yet to be determined. In this work, we show that Internet-based news sources can be used as the basis for a near real-time warning system for social response, especially for country-disease pairs with extensive Internet-based news media coverage. We used over 150,000 internet-based news articles provided by HealthMap, covering 16 diseases in 72 countries and territories around the globe, to validate our models' performance.

2 Methods

Our primary objective was to forecast social response in response to the spread of infectious disease. Our method consisted of three primary steps: (1) data acquisition and indicator extraction, (2) social response target development, and (3) social response forecasting. The data acquisition and indicator extraction step consisted of the collection of Internet-based news articles describing disease outbreaks around the world and automated tagging of these articles with indicators of the disease activity and social response. This process is described in Sect. 2.1. In Sect. 2.2 we explain how the social response indicator counts were translated into a target for prediction of future social response. This target was created by identifying periods of unusually severe social response based on the weekly aggregated social response indicator counts. The volume of Internet-based news reporting varies dramatically between countries and diseases, so the raw counts of the indicators were unsuitable for use as a target. Instead, we needed to create a target by comparing against baseline behavior. For example, in China 42% of weeks had at least one indicator of social response to avian influenza; 27% of weeks had over five indicators. Therefore, for China a couple mentions of social response to avian influenza per week may be considered normal behavior. In Zimbabwe, only 3% of weeks had one or more indicators of social response to cholera, making just one mention of social response an unusual event. We used a Bayesian network to compare each week's social response profile with a baseline for the country and disease. Then, we used a statistical process control algorithm to identify periods of time that were sufficiently unusual to be considered periods of social response. This approach was derived from approaches developed for rapid disease outbreak detection (Buckeridge et al. 2005; Wong et al. 2003). Outbreak detection algorithms take as an input syndromic surveillance data and output whether a disease outbreak is taking place. Our algorithm takes the social response indicator time series as an input and outputs whether an outbreak of social response is taking place. Finally, in Sect. 2.3 we describe the method developed for forecasting future social response. The entire approach is outlined in Fig. 1.

2.1 Data

HealthMap collects a continuous stream of near real-time information on disease outbreaks, including Internet-based news articles and government reports. Over 150,000 such free-text documents, collected between 2006 and 2015, were used for modeling. These documents

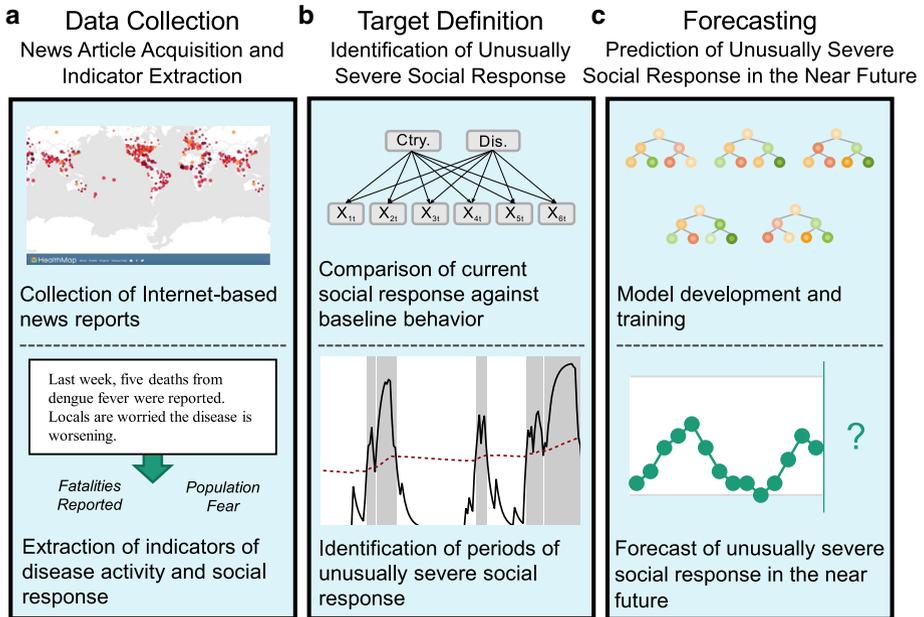


Fig. 1 Overview of methods. **a** First, news articles were automatically collected and tagged with indicators of disease activity and social response. **b** Next, an anomaly detection approach was used to identify periods of time with unusually severe social response profiles. These periods were used as targets for social response forecasting. **c** Finally, the occurrence of unusually severe social response was forecast for the coming week, 2 and 3 weeks

described breaking news events for 16 diseases¹ and 72 countries and territories.² The documents were automatically cleaned and, when necessary, translated into English.³ We have developed a natural language processing approach to automatically tag the documents with indicators describing the spread of the disease (4 indicators), the perceived severity of the disease (3 indicators), the preventative measures taken (7 indicators), and the social response (6 indicators; Affective Social Response: *Population Fear*, *Officials Fear*; Economic Social Response: *Economy Affected*, *Tourism Affected*; Behavioral Social Response: *Violence*, and *Healthcare Worker Protest*). These indicators were created by searching within each sentence of the text for combinations of words or phrases describing the events of interest. Eventually, these indicators could be expanded to not include current events of interest, but also events expected to occur in the future according to the news sources. The indicator counts were aggregated by week for each country and disease.

¹ The diseases were avian influenza, Chikungunya virus, cholera, dengue fever, influenza, Lassa fever, listeriosis, malaria, Marburg virus, measles, Middle East respiratory syndrome (MERS), norovirus, poliomyelitis (polio), West Nile virus, yellow fever, and Yersinia pestis (plague). These diseases were selected to include the most common diseases, as well as diseases with a number of different modes of transmission and severity levels.

² Countries were selected to achieve geographic and socio-economic diversity. There were 20 countries and territories included from the WHO African Region, 18 from the Americas Region, 9 from the Eastern Mediterranean Region, 11 from the European Region, 4 from the South-East Asia Region, and 10 from the Western Pacific Region.

³ HealthMap collects documents in 15 different languages.

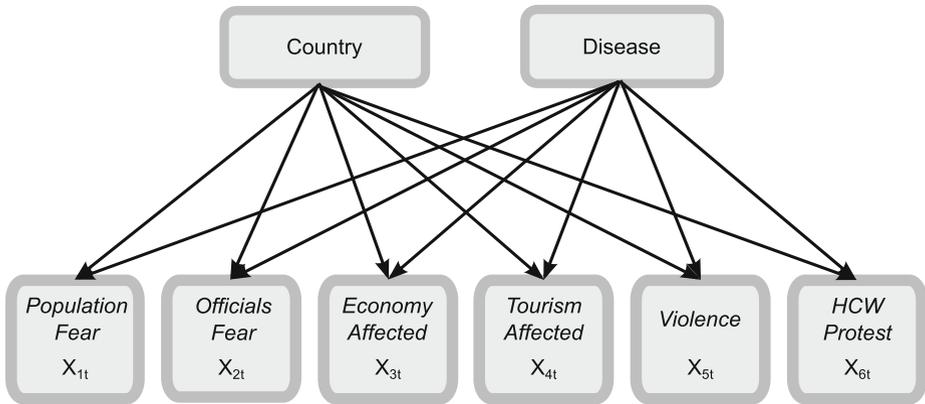


Fig. 2 Bayesian network describing relationships between country, disease, and social response indicator counts. All social response indicator counts were dependent upon the country and the disease. We allowed relationships between social response indicator counts (e.g. the count for *Violence* depends upon the count for *Population Fear*) to be learned, but did not require such relationships. The pictured network is the Bayesian network in the case where no relationships were learned among the social response indicator counts

2.2 Identifying unusually severe periods of social response

We used a Bayesian network to calculate the joint probability of a social response profile (the vector of social response indicator counts), given prior profiles for the same country and disease. Since Bayesian networks allow for aggregation of many types of signals, they are a popular method for anomaly detection (Buckeridge et al. 2005; Mascaro et al. 2014; Rashidi et al. 2011). In the developed Bayesian network, all social response indicator counts were dependent upon the country and disease. Dependencies in the network between social responses (e.g. the count for *Violence* depends upon the count for *Population Fear*) could be learned, but were not required to be present. The structure of the network was learned using a hill-climbing greedy search, with the Bayesian Information Criterion as the score. In order to train the network, we required that at least 2 years of news articles be collected for each country-disease pair. Figure 2 depicts the Bayesian network, with all required dependencies.

Let c_{ijkt} be the observed indicator count for social response indicator k^4 in country i for disease j during week t . Let x_{ijkt} be a discretized version of the indicator counts:

$$x_{ijkt} = \begin{cases} 0 & \text{if } c_{ijkt} = 0 \\ 1 & \text{if } c_{ijkt} = 1 \\ 2 & \text{if } 1 < c_{ijkt} \leq 5 \\ 3 & \text{if } 5 < c_{ijkt} \leq 20 \\ 4 & \text{otherwise.} \end{cases} \quad (1)$$

The splits used to discretize the social response indicator counts were selected empirically based on analysis of data. For 99.3% of weeks, no articles indicating *Population Fear* were collected; 0.6% of weeks had 1 article, 0.1% had between 2 and 5 articles, 0.02% had between 6 and 20 articles, and 0.003% had more than 20 articles. Similar patterns were observed for the other social response indicators.

⁴ Since six social response indicators were developed, $k \in \{1, 2, \dots, 6\}$.

Let X_{kt} be a random variable following the baseline distribution of social response indicator k , learned by the Bayesian network trained on weeks 1 through $t - 1$. Then, for each week t , country i , and disease j , we used likelihood weighting to calculate the probability of observing a social response profile as or more severe than the one observed during week t :

$$p_{ijt} = P(X_{1t} \geq x_{ij1t}, X_{2t} \geq x_{ij2t}, \dots, X_{6t} \geq x_{ij6t} \mid \text{Country} = i, \text{Disease} = j). \quad (2)$$

The probabilities were translated into anomaly scores (Mascaro et al. 2014):

$$A_{ijt} = -\ln(p_{ijt} + 0.01). \quad (3)$$

High anomaly scores indicate weeks with abnormally severe social response profiles, compared with previous weeks. For example, a week with a probability of 5% would have an anomaly score of 2.8. A week with a probability of 80% would have an anomaly score of 0.2.

The next step was to identify multi-week periods of unusually severe social response, using the weekly anomaly scores, A_{ijt} . For this task, we used the exponentially weighted moving average (EWMA) (Roberts 1959) of the anomaly scores. Alternative approaches to finding statistical breakpoints in social media data have been proposed (Servi 2013). Nevertheless, researchers have found that EWMA is a “simple and robust” method for outbreak identification based on surveillance of sparse syndromic data (Buckeridge et al. 2005), and, continuing the analogy of social response to disease, it is reasonable to expect that EWMA would provide good performance on a sparse data stream of social response anomaly scores. The EWMA, Z_{ijt} , is the weighted average of all previous anomaly scores and is defined as follows:

$$Z_{ijt} = \lambda A_{ijt} + (1 - \lambda) Z_{ijt-1}, \quad (4)$$

where $\lambda \in (0, 1)$. Since 2 years (104 weeks) of news articles were collected before the anomaly scores were calculated, the EWMA was started on the 105th week, and $Z_{ij104} = 0$. We defined a binary indicator for the presence of unusually severe social response, which was 1 when the EWMA of the anomaly scores exceeded the upper control limit (UCL_{ijt}) and 0 otherwise:

$$S_{ijt} = \begin{cases} 1 & \text{if } Z_{ijt} > UCL_{ijt} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

In Sect. 2.3, we introduce models to predict the probability that $S_{ijt} = 1$ in the coming 1, 2, and 3 weeks. The upper control limit for an EWMA control chart is defined as follows (Montgomery 2009):

$$UCL_{ijt} = \mu_{ijt} + L\sigma_{ijt}\sqrt{\frac{\lambda}{2-\lambda}(1 - (1-\lambda)^{2(t-104)})}, \quad (6)$$

with width of the control limit, $L > 0$, in-control mean, $\mu_{ijt} = \frac{1}{t-105} \sum_{v=105}^{t-1} Z_{ijv}$, and in-control standard deviation, $\sigma_{ijt} = \sqrt{\frac{1}{t-105} \sum_{v=105}^{t-1} (Z_{ijv} - \mu_{ijt})^2}$.

In the standard implementation of EWMA, both the upper control limit and EWMA are reset after the EWMA passes the control limit. We found that S_{ijt} was most reasonable when these values were not reset. The EWMA parameter, L , was set to 3 based on the recommendation of Montgomery (2009). The parameter, λ , was tuned by visually inspecting the S_{ijt} indicators for several values. The tuning process used data from 30 countries. Prediction results for these countries are presented in Online Resource 2. The selected value, $\lambda = 0.25$, produced results for S_{ijt} that corresponded well with analyst opinion. Figure 3 shows the

social response indicator counts, the exponentially weighted moving average, and the social response binary indicator for dengue fever outbreaks in India. Figures depicting several other countries and diseases can be found in Online Resource 1.

Our approach to defining the binary social response indicator has a number of advantages. First, especially for country-disease pairs with high volumes of Internet-based media attention, the social response indicator is robust to errors in the automatic tagging of the social response indicators. A single incorrect indicator will typically be insufficient to produce an anomaly score that is high enough to cause the EWMA to cross the control limit. While data cleaning could be used to limit the effect of incorrect indicators, it also risks accidental removal of true indicators. We believe that EWMA is a more conservative approach, and is more suitable to our particular problem. A second advantage is that the social response indicator is comparable across countries and diseases, since it is defined relative to a baseline for the country and disease, removing the effect of differing volumes of media coverage. Finally, the social response indicator is interpretable. A value of 1 always indicates that an unusually severe social response signal has been observed.

2.3 Prediction of future social response

Now, we introduce the approach to forecasting unusually severe social response the coming 1, 2, and 3 weeks (see Fig. 1c). The news articles were transformed and structured into time-series, cross-section data with a binary dependent variable (BTSCS). This type of data structure has been previously studied (Beck et al. 1998), with the key observation that BTSCS data are grouped duration data. Therefore, it is essential to predict the timing of (1) the transition from a state of no social response to a state of social response, and (2) the transition from a state of social response to a state of no social response. Note that transition from a state of no social response to a state of social response is a rare event, and most frequently our models predict that no transition will take place. Also, note that the signal indicating a transition from a state of no social response into a state of social response is likely different from the signal indicating a continuation of social response once it has already begun. It has been suggested that separate models should be built to predict the transitions into and out of a binary state (Beck et al. 2001; Jackman 2000), and we adopted that suggestion here for our binary social response indicator.

2.3.1 Target definition

Because we were interested in forecasting social response over time horizons longer than 1 week, we defined a target, Y_{ijt}^w , indicating the occurrence of social response for disease j in country i in the w weeks following week t :

$$Y_{ijt}^w = \begin{cases} 1 & \text{if } S_{ij(t+1)} + \dots + S_{ij(t+w)} \geq 1 \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Note that $Y_{ijt}^1 = S_{ij(t+1)}$, but in general Y_{ijt}^w is not equivalent to $S_{ij(t+w)}$.

2.3.2 Transition models

We built two models: Model $0 \rightarrow 1$ and Model $1 \rightarrow 0$. Model $0 \rightarrow 1$ predicted the transition from a period of no social response into a period of social response [i.e. Model $0 \rightarrow 1$ estimates $P(Y_{ijt}^w = 1 | S_{ijt} = 0)$], and Model $1 \rightarrow 0$ predicted whether the period of

Table 1 Features of the social response transition models

Model 0 → 1 features	Model 1 → 0 features
Country name	Country name
Disease name	Disease name
Number of articles about country and disease in week (with and without duplicates)	Number of articles about country and disease in week (with and without duplicates)
Indicator counts for all 20 indicators	Indicator counts for all 20 indicators
3 week slopes of indicator counts	EWMA as a percentage of UCL $\left(\frac{Z_{ijt}}{UCL_{ijt}}\right)$
Duration of current no social response period	3 week slope of EWMA
	Duration of current social response period

social response would continue or instead transition back into a period of no social response [i.e. Model 1 → 0 estimates $P(Y_{ijt}^w = 1 \mid S_{ijt} = 1)$]. A subset of the data, consisting of 30 countries was used for parameter tuning and feature selection. We show results separately for this subset in Online Resource 2. The features included in the final models are listed in Table 1.

During periods with no social response ($S_{ijt} = 0$), we used Model 0 → 1 to anticipate the onset of social response. Following social response onset ($S_{ijt} = 1$), we used Model 1 → 0 to predict the end of social response. The Model 0 → 1 training set consisted of all observations from weeks 105 through $t - w$. The Model 1 → 0 training set consisted only of observations from weeks 105 through $t - w$ that occurred within a period of social response (i.e. the observation from country i , disease j , and time t_0 would be included if $t_0 \leq t - w$ and $S_{ijt_0} = 1$).⁵ The training data was kept separate from the testing data by training the model on weeks 105 through $t - w$ and testing on week t for all $t > 105$.

Since transitions from periods of no social response to periods of social response were extremely rare events,⁶ the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002) was used on the Model 0 → 1 training set to increase the prevalence of the target to 20%. Edited nearest-neighbors (ENN) (Wilson 1972) was then used to remove examples that were misclassified by two of three nearest-neighbors. The combination of SMOTE and ENN has been shown to be effective for a number of prediction problems involving imbalanced data (Batista et al. 2004). For both Model 0 → 1 and Model 1 → 0, features that had near-zero variance in the training data were removed. Finally, a random forest with 100 trees was trained (Breiman 2001), and a prediction was generated for Y_{ijt}^w . Since the sequence of features is of interest in this problem, Hidden Markov Models could be considered as an alternative classifier (Rabiner and Juang 1986).

3 Results

The model performance was evaluated on historical data for each of three time horizons: next week, next 2 weeks, and next 3 weeks. For each country-disease pair, 2 years of training

⁵ The training period began at week 105, since 2 years, 104 weeks, of data were used to calibrate the anomaly scores.

⁶ In our data, 826 transitions from periods of no social response to periods of social response were observed. There were 222,428 weeks with no social response, resulting in 0.4% prevalence of the target Y_{ijt}^1 .

Table 2 Overall performance of the social response prediction models. Model performance was evaluated based on six metrics: accuracy, sensitivity, sensitivity looking only at weeks with articles in the preceding 3 weeks, sensitivity looking only at weeks with articles in the preceding week, specificity, and precision. Model 0 → 1 predicted the onset of periods of social response, while Model 1 → 0 predicted the end of such periods

	Accuracy	Sensitivity			Specificity	Precision
		All weeks	Weeks with one or more articles in prior 3 weeks	Weeks with one or more articles in prior week		
Next week (Y_{ijt}^1)						
Model 0 → 1	0.991	0.213	0.389	0.460	0.995	0.146
Model 1 → 0	0.923	0.956	–	–	0.821	0.943
Next 2 weeks (Y_{ijt}^2)						
Model 0 → 1	0.988	0.179	0.367	0.452	0.994	0.193
Model 1 → 0	0.916	0.954	–	–	0.780	0.940
Next 3 weeks (Y_{ijt}^3)						
Model 0 → 1	0.984	0.162	0.363	0.457	0.993	0.221
Model 1 → 0	0.909	0.951	–	–	0.745	0.936

data were observed before the first target was predicted for model performance evaluation. In the results, we show how performance is affected by the length of the prediction window (1, 2, or 3 weeks) and by the volume of news articles published for the country-disease pair. Online Resource 2 provides additional summarized prediction results, including results for the model features and results for the set of 30 countries that were used for initial model construction and tuning.

3.1 Model performance metrics

We used several metrics to evaluate the performance of our model: accuracy, sensitivity, specificity, and precision.⁷ In addition, we evaluated models’ sensitivity looking only at weeks with social response that had at least one news article published in the prior one or prior 3 weeks. The two additional sensitivity metrics were used because a large percentage of weeks with social response, 48%, had no articles on the disease in the preceding 3 weeks. Because there were no articles in the preceding weeks, those targets were essentially impossible to predict using data from news articles. Therefore, we wanted to assess our model’s sensitivity excluding such weeks.

3.2 Model performance

The developed models achieved good performance for country-disease pairs with substantial media coverage, and fair performance for country-disease pairs with little coverage. Table 2 shows both Model 0 → 1 and Model 1 → 0 performance aggregated for all country-

⁷ Let TP be the number of true positive predictions, TN be the number of true negative predictions, FP be the number of false positive predictions, and FN be the number of false negative predictions. Then, Accuracy = $(TP + TN)/(TP + FP + TN + FN)$, Sensitivity = $TP/(TP + FN)$, Specificity = $TN/(TN + FP)$, and Precision = $TP/(TP + FP)$.

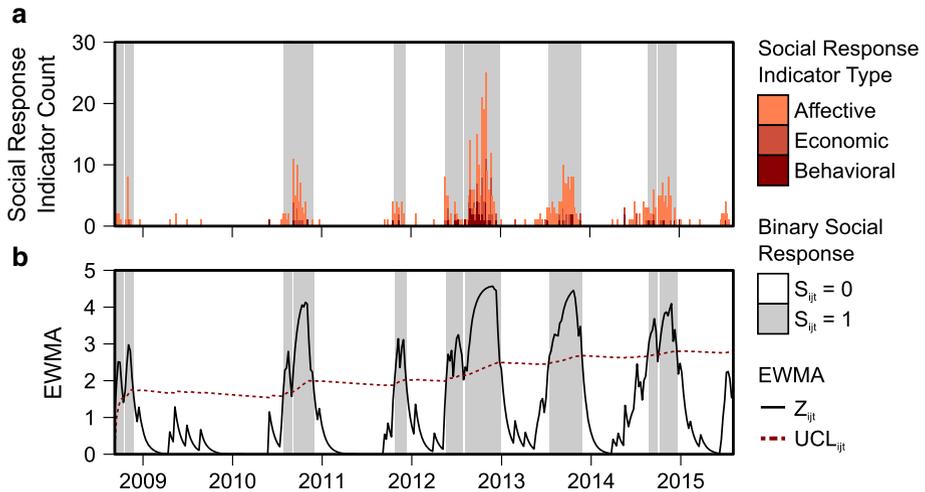


Fig. 3 Identification of periods of unusually severe social response for dengue fever outbreaks in India. **a** The social response indicator counts are shown by social response type. Overall, the peaks in social response indicator counts align well with the binary social response indicator, S_{ijt} . **b** The exponentially weighted moving average of the anomaly scores (Z_{ijt}) is shown along with the upper control limit (UCL_{ijt}). The binary social response indicator is 1 when the exponentially weighted moving average surpasses the control limit

disease pairs. Model $0 \rightarrow 1$ exhibited 46% sensitivity in predicting the onset of a social response period in the next week for weeks with at least one articles in the preceding 3 weeks. Model predictions over longer time horizons were slightly less sensitive, but substantially more precise. Model $0 \rightarrow 1$'s relatively low precision for the target Y_{ijt}^1 appears to result largely from premature prediction of social response. Twenty-four percent of Model $0 \rightarrow 1$ false positive predictions for Y_{ijt}^1 occurred in the 6 weeks prior to the onset of a period of social response. In these cases, the model likely detected indications that the situation was worsening, but predicted that the transition into a period of social response would take place sooner than actually occurred. Model $1 \rightarrow 0$ consistently predicted the end of periods of social response for all time horizons, with over 74% specificity and over 90% accuracy.

Figure 4 shows the predictions for social response in the next 2 weeks for dengue fever outbreaks in India. During no social response periods, the model predicted a low probability of social response in the next 2 weeks. As the onset of a period of social response was approached, the predicted probability of social response increased. As the period of social response ended, the predicted probabilities fell. Additional figures depicting results for other countries and diseases can be found in Online Resource 3.

The performance of the model varied depending upon the quantity of Internet-based news reporting for the country-disease pair. Table 3 compares the Model $0 \rightarrow 1$ performance for country-disease pairs that had a median of 20 or more articles per year in our data⁸ with performance for pairs that had median of fewer than 20 articles per year. The model performance was greatly improved with higher media volume. For country-disease pairs

⁸ There were 25 such country-disease pairs: measles in Australia; influenza, measles, and norovirus in Canada; avian influenza, dengue fever, influenza, and measles in China; cholera in Cuba, avian influenza in Egypt; norovirus in the United Kingdom; dengue fever in Honduras; dengue fever, influenza, and malaria in India; norovirus in Japan; dengue fever in Malaysia; cholera and polio in Nigeria; dengue fever and polio in Pakistan; dengue fever in Peru; dengue fever in the Philippines; and avian influenza and dengue fever in Vietnam.

Table 3 Comparison of performance for predicting the onset of social response (Model 0 → 1) for country-disease pairs with a median of 20 or more news articles per year and those with fewer articles per year. Model performance was evaluated based on six metrics: accuracy, sensitivity, sensitivity looking only at weeks with articles in the preceding 3 weeks, sensitivity looking only at weeks with articles in the preceding week, specificity, and precision. Model sensitivity and precision were dramatically higher for the country-disease pairs with a median of 20 or more articles per year, than for the pairs with fewer articles per year

	Accuracy	Sensitivity			Specificity	Precision
		All weeks	Weeks with one or more articles in prior 3 weeks	Weeks with one or more articles in prior week		
Next week (Y_{ijt}^1)						
≥20 articles	0.887	0.602	0.633	0.671	0.897	0.169
<20 articles	0.994	0.123	0.260	0.311	0.997	0.127
Next 2 weeks (Y_{ijt}^2)						
≥20 articles	0.858	0.536	0.569	0.619	0.877	0.206
<20 articles	0.991	0.107	0.265	0.337	0.997	0.182
Next 3 weeks (Y_{ijt}^3)						
≥20 articles	0.838	0.516	0.552	0.607	0.864	0.238
<20 articles	0.988	0.096	0.267	0.348	0.996	0.206

with a median of 20 or more articles per year, the onset of social response in the next week was correctly predicted over 60% of the time (67% of the time among events with articles in the past week). The overall accuracy of the model was over 83% for each of the three time horizons. High accuracy (over 98% for all three time horizons) was achieved for country-disease pairs with a median of less than 20 articles per year, but the model was not successful at predicting the onset of social response, with only 12% sensitivity for Model 0 → 1 in predicting the occurrence of social response in the next week. Sensitivity was much higher, 31%, when looking only at weeks with one or more articles published in the prior week, suggesting that lack of articles in the weeks preceding the onset of social response contributes to the low sensitivity of Model 0 → 1 for country-disease pairs with median news coverage below 20 articles per week. Model 1 → 0 performance for different volumes of news media coverage is shown in Online Resource 3.

4 Discussion

The presented results confirm that information derived from Internet-based news sources not only provides early and accurate information for disease detection and analysis of spread, but can also be successfully used for detecting and predicting social response associated with detected disease. The developed models predicting the onset of social response and monitoring its progress and subsequent decline achieved good performance for diseases that receive substantial media attention in the country in which they are spreading. For country-disease pairs with a median of more than 20 articles per year in our data, the onset of social response in the next week was correctly predicted over 60% of the time. Sensitivity was higher still, 67%, when looking only at social response events with news articles published in the prior week. The continuation of periods of social response was predicted with over

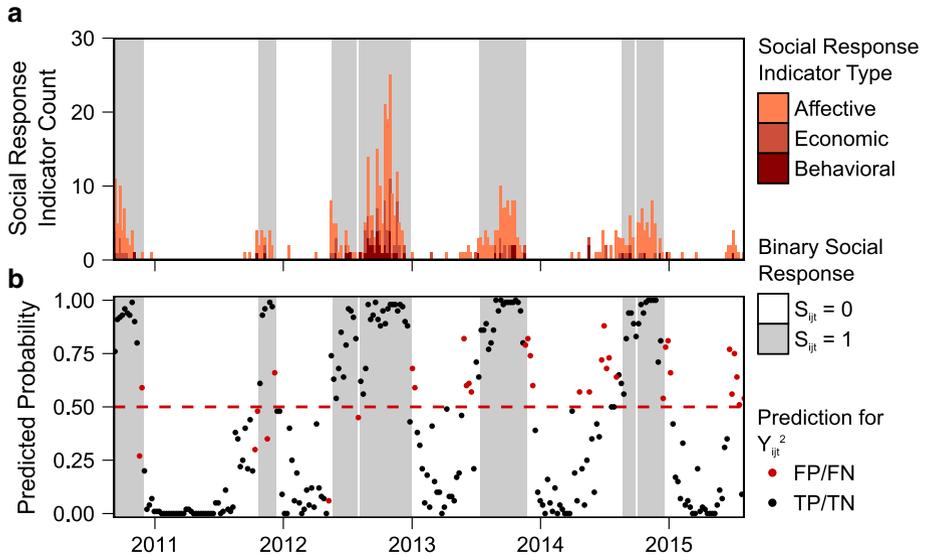


Fig. 4 Predicted probability of unusually severe social response in the next 2 weeks for dengue fever outbreaks in India. **a** The social response indicator counts are shown by social response type. Periods during which social response was occurring ($S_{ijt} = 1$) are shaded in grey. **b** The predicted probability of social response in the next 2 weeks ($P(Y_{ijt}^2 = 1)$) is shown. The predictions are colored according to whether an incorrect (false positive or false negative; red) or correct (true positive or true negative; black) prediction was made. Overall, the predictions exhibit the desired behavior—low probability of social response in the next 2 weeks was predicted during no social response periods, and, as a social response period were approached, the predicted probability increased. (Color figure online)

95% success for all time horizons. Their end was also predicted consistently, with over 74% success for all time horizons. Compared with predictions for social response in the coming week, predictions for social response in the coming 2 and 3 weeks were slightly less sensitive (39% for the next 3 weeks vs. 36% for the next week), but substantially more precise (22% for the next 3 weeks vs. 15% for the next week). Thus, in practice, predictions over relatively long time horizons may be most useful.

Country-disease pairs that received little media attention were not good candidates for predicting the onset of future social response. Internet-based news reporting on these pairs often does not begin until after social response has already started. For country-disease pairs with a median of less than 20 articles per year, 58% of weeks with social response had no articles about the disease in the prior 3 weeks; 85% had three or fewer articles. For such diseases, the main utility of our system is in identifying social response when it occurs, rather than predicting when it will happen in the future. There are several reasons why a disease would receive little media attention in a country prior to an outbreak. One reason is that the country has a relatively undeveloped online news reporting system, and few articles are published about any type of disease transmission. Other possible reasons are that the disease is perceived as benign and not newsworthy, or that government censorship suppresses reporting. In these cases, it is possible that alternative data sources⁹ could be used to supplement data from Internet-based news media to improve prediction of future social

⁹ Examples of possible data sources include Internet search engine interest in the disease, social media posts, sales data for medical supplies, satellite images, radio broadcasts, and climate data.

response. Another reason why a disease would receive little reporting prior to an outbreak is that the disease is newly introduced into the country. In our data, the emergence of a new disease in a country is frequently associated with social response. There is little that can be done to improve prediction of the onset of this social response, because forecasting the exact timing of the introduction of a disease into a country is beyond the ability of current biosurveillance techniques.

In summary, we have developed an approach for anticipating social response to infectious disease spread in near real-time, and have evaluated it using outbreaks of 16 different diseases in 72 locations around the world. We have demonstrated that Internet-based news can serve as a good data source for predicting social reaction to disease spread, when there is sufficient news coverage of the disease. In general, our system is most effective for countries with active Internet-news reporting systems and for diseases that receive frequent coverage—avian influenza, cholera, dengue fever, influenza, malaria, measles, and polio. By identifying ongoing social response and alerting decision makers and biosurveillance experts to probable social response in the near future, this warning system will provide responders with the information needed to better combat both the disease spread itself and its detrimental social consequences.

References

- Batista, G. E. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. doi:10.1145/1007730.1007735.
- Beck, N., Epstein, D., Jackman, S., & O'Halloran, S. (2001). Alternative models of dynamics in binary time-series-cross-section models: The example of state failure. <http://hdl.handle.net/10022/AC:P:9718>.
- Beck, N., Katz, J. N., & Tucker, R. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science*, 42(4), 1260–1288. doi:10.2307/2991857.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med*, 5(7), e151. doi:10.1371/journal.pmed.0050151.
- Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., & Moore, A. W. (2005). Algorithms for rapid outbreak detection: A research synthesis. *Journal of Biomedical Informatics*, 38(2), 99–113.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1), 321–357.
- Cheng, C. (2004). To be paranoid is the standard? Panic responses to SARS outbreak in the Hong Kong Special Administrative Region. *Asian Perspective*, 28(1), 67–98.
- Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., et al. (2008). Biocaster: Detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24), 2940–2941. doi:10.1093/bioinformatics/btn534.
- D'Orazio, V., & Yonamine, J. E. (2015). Kickoff to conflict: A sequence analysis of intra-state conflict-preceding event structures. *PLoS ONE*, 10(5), e0122472. doi:10.1371/journal.pone.0122472.
- Doyle, A., Katz, G., Summers, K., Ackermann, C., Zavorin, I., Lim, Z., et al. (2014). Forecasting significant societal events using the EMBERS streaming predictive analytics system. *Big Data*, 2(4), 185–195. doi:10.1089/big.2014.0046.
- Fast, S. M., González, M. C., Wilson, J. M., & Markuzon, N. (2015). Modelling the propagation of social response during a disease outbreak. *Journal of The Royal Society Interface*, 12(104), 20141105. doi:10.1098/rsif.2014.1105.
- Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, 31(6), 649–679.

- Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115–125.
- International Federation of Red Cross and Red Crescent Societies (2015) Red Cross Red Crescent denounces continued violence against volunteers working to stop the spread of Ebola. <http://www.ifrc.org/en/news-and-media/press-releases/africa/guinea/red-cross-denounces-continued-violence-against-volunteers-working-to-stop-the-spread-of-ebola>
- Jackman, S. (2000). In and out of war and peace: Transitional models of international conflict. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.200.5895&rank=1>.
- Kinsman, J. (2012). “A time of fear”: Local, national, and international responses to a large Ebola outbreak in Uganda. *Globalization and Health*, 8, 15–15.
- Lau, J. T. F., Griffiths, S., Choi, K. C., & Tsui, H. Y. (2010). Avoidance behaviors and negative psychological responses in the general population in the initial stage of the H1N1 pandemic in Hong Kong. *BMC Infectious Diseases*, 10(1), 139. doi:10.1186/1471-2334-10-139.
- Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the global burden of disease study 2010. *Lancet*, 380(9859), 2095–2128. doi:10.1016/S0140-6736(12)61728-0.
- Mascaro, S., Nicholso, A. E., & Korb, K. B. (2014). Anomaly detection in vessel tracks using Bayesian networks. *International Journal of Approximate Reasoning*, 55(1), 84–98.
- McGrath, J. W. (1991). Biological impact of social disruption resulting from epidemic disease. *American Journal of Physical Anthropology*, 84(4), 407–419. doi:10.1002/ajpa.1330840405.
- Montgomery, D. C. (2009). *Introduction to Statistical Quality Control* (6th ed.). New Jersey: Wiley.
- Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis*, 20(3), 271–291.
- Mykhalovskiy, E., & Weir, L. (2006). The global public health intelligence network and early warning outbreak detection: A Canadian contribution to global public health. *Canadian Journal of Public Health/Revue Canadienne de Santé Publique*, 97(1), 42–44.
- O’Brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1), 87–104. doi:10.1111/j.1468-2486.2009.00914.x.
- Rabiner, L. R., & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1), 4–16.
- Racette, M. P., Smith, C. T., Cunningham, M. P., Heekin, T. A., Lemley, J. P., & Mathieu, R. S. (2014). Improving situational awareness for humanitarian logistics through predictive modeling. *Systems and Information Engineering Design Symposium (SIEDS), 2014*, 334–339.
- Rashidi, L., Hashemi, S., & Hamzeh, A. (2011). Anomaly detection in categorical datasets using Bayesian networks. *Artificial Intelligence and Computational Intelligence*, 7003, 610–619.
- Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3), 239–250.
- Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2), 1–19. doi:10.1145/1462198.1462204.
- Servi, L. (2013). Analyzing social media data having discontinuous underlying dynamics. *Operations Research Letters*, 41(6), 581–585.
- Sherlaw, W., & Raude, J. (2013). Why the French did not choose to panic: A dynamic analysis of the public response to the influenza pandemic. *Sociology of Health & Illness*, 35(2), 332–344.
- Truvé, S. (2013). Big data for the future: Unlocking the predictive power of the web. <http://www.slideshare.net/RecordedFuture/big-data-for-the-future-unlocking-the-predictive-power-of-the-web>
- Vaisman, E., Fast, S. M., Cunha, M. G., Postlethwaite, T., Wilson, J. M., & Mekaru, S. R. (2014). Predicting negative social response to disease outbreaks using biosurveillance and news data. In: 2014 INFORMS Workshop on Data Mining and Analytics.
- Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, 2(3), 408–421.
- Wilson, K., & Brownstein, J. S. (2009). Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal*, 180(8), 829–831. doi:10.1503/cmaj.1090215.
- Wong, W.K., Moore, A., Cooper, G., & Wagner, M. (2003). Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 808–815).
- Woodall, J. P. (2001). Global surveillance of emerging diseases: The ProMED-mail perspective. *Cad Saude Publica*, 17(Suppl), 147–154.