PUBLIC HEALTH

Fine-grained dengue forecasting using telephone triage services

10.1126/sciadv.1501215

Nabeel Abdur Rehman,[1,2] Shankar Kalyanaraman,[3,4] Talal Ahmad,[3,4] Fahad Pervaiz,[5] Umar Saif,[1,6] Lakshminarayanan Subramanian[3,4]*

Thousands of lives are lost every year in developing countries for failing to detect epidemics early because of the lack of real-time disease surveillance data. We present results from a large-scale deployment of a telephone triage service as a basis for dengue forecasting in Pakistan. Our system uses statistical analysis of dengue-related phone calls to accurately forecast suspected dengue cases 2 to 3 weeks ahead of time at a subcity level (correlation of up to 0.93). Our system has been operational at scale in Pakistan for the past 3 years and has received more than 300,000 phone calls. The predictions from our system are widely disseminated to public health officials and form a critical part of active government strategies for dengue containment. Our work is the first to demonstrate, with significant empirical evidence, that an accurate, location-specific disease forecasting system can be built using analysis of call volume data from a public health hotline.

INTRODUCTION

The province of Punjab in Pakistan, with a population of close to 100 million (*1*), was affected by a dengue epidemic in 2011. More than 21,000 dengue patients were reported, most of them from the city of Lahore (*2*). There was no mechanism to detect the epidemic early or localize the outbreaks, resulting in the loss of more than 350 lives.

Because there is no known cure or vaccine for treating different stages of dengue fever (*3*, *4*), most public health efforts focus on prevention through a combination of active disease surveillance and vector control methods (*5*, *6*). These methods target source reduction to eliminate the breeding grounds of the vector through environmental waste management, water management, and biological and chemical measures. The World Health Organization advocates the use of early warning systems to signal outbreaks ahead of time to contain diseases such as dengue.

To develop an early warning system, real-time surveillance data are required. In developing countries, such as Pakistan, conventional health data gathering methods are error-prone and take weeks to compile. This leads to a delay in containment response against a disease, resulting in a pandemic or seasonal outbreak. In particular, accurate forecasting of the number of future patients and their location gives government agencies adequate time to mobilize and target resources for containment and spread awareness (*7*, *8*). These early containment efforts curb the spread of disease at an early stage, restricting it from turning into a citywide outbreak and affecting a large population.

Epidemiological data and geographical information system data have previously been used for fine-grained disease surveillance (*9*); however, the data-intensive nature of these systems requires the availability of a large workforce for data collection, often not available in the developing world. Besides the cost of these systems, collection of disease data requires significant training and regulatory interventions to ensure that all health facilities, in both public and private sectors, report accurate disease data in real time.

Therefore, in resource-constrained developing countries, it may only be possible to estimate the spread of disease using indirect methods. Previous works have explored the use of telephone triage services (*10*), Internet search queries (*11*–*13*), online media reports (*14*), and environmental parameters (*15*) as indirect data sources to build epidemic warning systems. However, systems based on environmental parameters and online media reports provide only coarse-grained surveillance because of the nature of the data sources, whereas Internet search query–based systems only perform well in countries with high Internet penetration. Data from telephone triage services (*10*) and health hotline facilities (*16*) have been shown to have a correlation with influenza activity; NHS Direct (*17*) uses calls to a health hotline to generate alerts to complement other surveillance methods. However, previous studies have reported that telephone triage services are not a reliable source of surveillance data on a national scale because of variability in coverage and a lack of statistical prediction models from these data (*10*).

Here, we present results of how we used a simple phone-based helpline facility (telephone triage service) to develop an early epidemic warning system for dengue in Pakistan. Contrary to previous experience, our system provides an accurate measure of future disease cases at a fine-grained subcity level. Our system not only flags an outbreak (*17*) but also makes an accurate forecast (median correlation of 0.85) of both the number of patients and their locations 2 to 3 weeks ahead of time. The ability of our system to accurately forecast patients and their locations is critical for the government to mobilize and target its resources to contain an outbreak. Our prediction model carefully incorporates both weather indicators and awareness level in a community to reduce the false alerts common in previous crowd-sourced systems (*18*).

We report on our experiences running the system since the start of 2012 in Lahore, Pakistan (see Materials and Methods). With more than 300,000 hotline calls, we provide block-level forecasts of dengue cases to health organizations throughout the city and constantly validate our inferences from case data gathered from hospitals in the city. The appeal of our model is its usefulness despite its sheer simplicity; to our best knowledge, our system is the first to demonstrate, with significant empirical evidence, that an accurate, fine-grained, and locality-specific disease forecasting system can be built using analysis of call volume data from a public health hotline (or a telephone triage service) (*19*, *20*).

[1]Information Technology University, Lahore 54000, Pakistan. [2]Computer Science and Engineering, New York University, New York, NY 11201, USA. [3]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA. [4]Center for Technology and Economic Development, NYU Abu Dhabi, Abu Dhabi PO Box 129188, United Arab Emirates. [5]Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA. [6]Punjab Information Technology Board, Lahore 54000, Pakistan.
*Corresponding author. Email: lakshmi@cs.nyu.edu

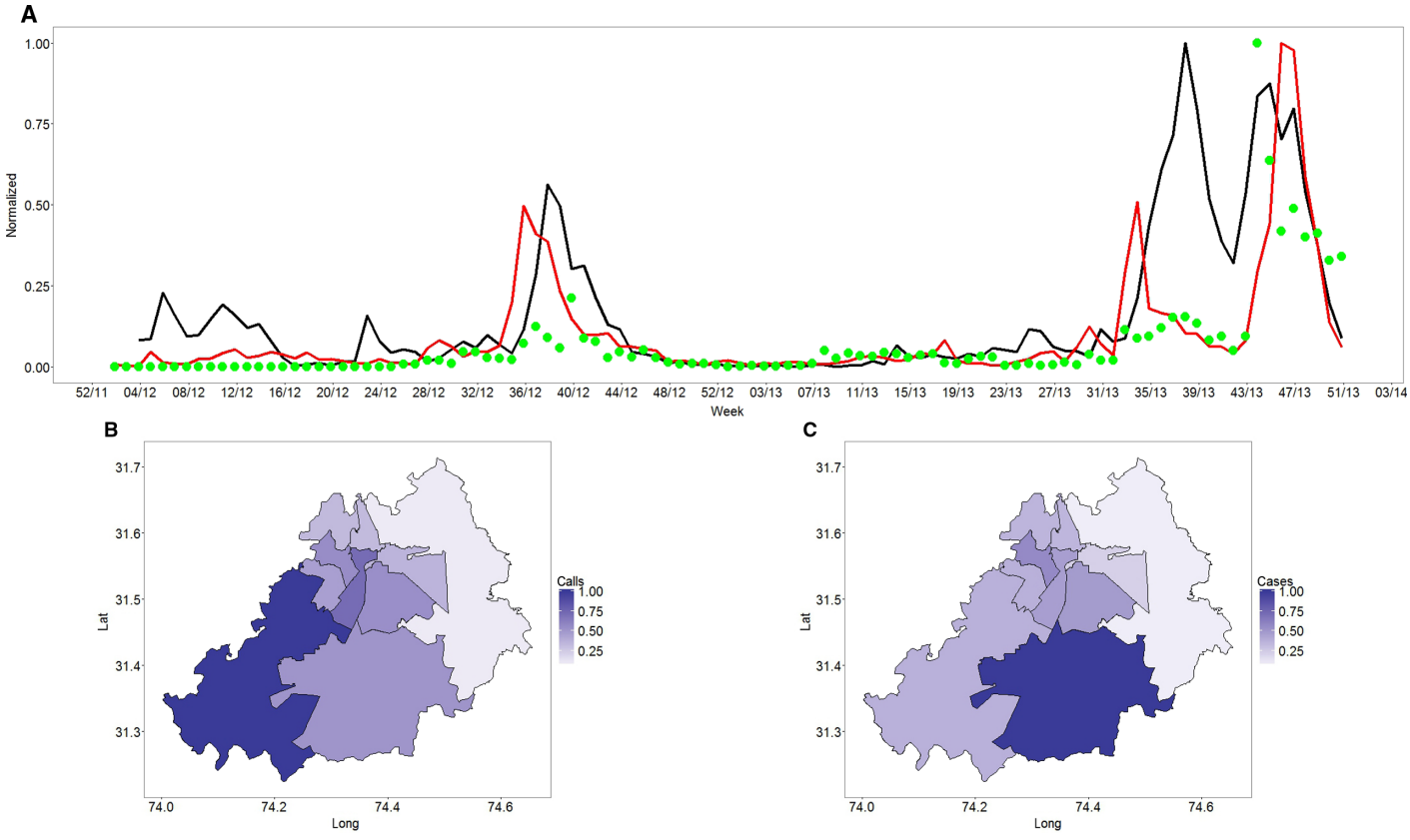

Fig. 1. Trends in call volume and suspected dengue cases measured during 2012 and 2013. (**A**) Time series of calls (red), suspected dengue cases (black), and awareness campaigns (green points). Scale normalized by dividing by individual maximum values. The *x*-axis label is in week of the year. (**B**) Density map of calls across towns in Lahore. (**C**) Density map of cases across towns in Lahore. The lightest shade represents the least number, and the darkest shade represents the highest number. The legend is normalized by the maximum value. Lat, latitude; long, longitude.

RESULTS

Figure 1 shows the baseline data used in our analysis. Figure 1A highlights the high correlation between the number of dengue patients reported in hospitals and the raw number of "dengue symptom inquiry" calls received at the health hotline in Lahore for the years 2012–2013. In the data shown in Fig. 1A, the number of calls "lead" the number of patients. Although the number of calls aggregated at the city level has a strong correlation with the number of cases during the year 2012 as shown in fig. S1, we observe that their relative variations across time and across different towns were not uniform. Hence, calls cannot be used alone to predict the number of cases and additional parameters, such as awareness level and climate conditions needed to be incorporated in the model. Figure S1 shows the cross-correlation values between suspected cases reported from hospitals and the number of calls received at the health hotline for the city of Lahore during 2012. Figure 1 (B and C) shows that 7 of 10 towns in Lahore follow a similar pattern: the relative amount of calls received from a town exhibits similar variations to the relative number of patients reported from the same town.

Although our baseline analysis shows a correlation between call volume data and disease cases, the calling patterns of citizens are naturally dependent on the level of awareness about the telephone triage service. This is visible in Fig. 1A. During the first peak of cases in 2013, the awareness campaign activities are low, resulting in a lower number of calls being received at the health hotline. In contrast, during the second peak of cases in 2013, the awareness campaign activities are high, resulting in a higher number of hotline calls. Previous studies that used data from alternative data sources did not incorporate this aspect; however, for public hotline services, it must be explicitly incorporated to address the inaccuracies introduced by variable levels of awareness across towns. Furthermore, previous studies that reported weak results (*10*) from the use of triage service for epidemic detection only relied on call volume data and ignored other factors, such as weather conditions. Because temperature, precipitation, and humidity are critical to mosquito survival, reproduction, and development and can influence mosquito presence and abundance, our model also incorporates city-level weather data (*21*).

Our dengue forecasting system is based on an ensemble model and uses the random forest learning algorithm. We predict $\log(S(w+2,l))$ using $C(w,l)$, $A(w,l)$, $H(w)$, $T(w)$, $R(w)$, where $S(w+2,l)$ is the suspected cases reported during the second week after week w from town l; $C(w,l)$ and $A(w,l)$ are the number of calls received by health hotline and the number of awareness campaigns carried out in week w in town l, respectively; and $H(w)$, $T(w)$, and $R(w)$ are the average humidity, average temperature, and average rainfall during week w in the city (see Materials and Methods). The estimated model showed that the number of calls was the most important feature to forecast log-suspected cases, followed by average temperature (see Table 1). The estimates generated obtained a good fit with log-suspected cases reported from the government hospitals, with a median root mean square error (RMSE) of 0.80 (minimum, 0.52; maximum, 1.1; n = 10 towns) and a median correlation of 0.85 (minimum, 0.80; maximum, 0.93; n = 10 towns). Figure 2 shows the

Table 1. Random forest importance weights for parameters of the model trained over the total year and season (July to November).

Parameter	Average IncNodePurity	
	Total	Season
Calls	424.34	172.08
Awareness	274.28	158.41
Rainfall	123.14	50.89
Humidity	287.40	81.65
Temperature	349.38	137.61

model forecasts for 2012 and 2013 across all the towns in Lahore. In Pakistan, towns are the second smallest administrative units. Towns in Lahore have a median area of 57 km^2 (minimum, 24; maximum, 516; n = 10 towns).

To highlight the value added to the current model by the addition of data about the number of patients, the log($S(w,l)$) term was introduced in the predictors. A similar methodology was used to estimate the dengue cases 2 weeks in the future (see Materials and Methods). The estimates generated from the new model obtained a median RMSE of 0.63 (minimum, 0.50; maximum, 0.82; n = 10 towns) and a median correlation of 0.88 (minimum, 0.80; maximum, 0.94; n = 10 towns), with log-suspected cases reported from the government hospitals as shown in Fig. 3.

Estimates for dengue cases 3 weeks in the future, denoted log($S(w+3,l)$), were also generated using a similar methodology. Estimates from the model without data about the number of patients as a predictor obtained a median RMSE of 0.80 (minimum, 0.54; maximum, 1.1; n = 10 towns) and a median correlation of 0.84 (minimum, 0.77; maximum, 0.91; n = 10 towns), whereas estimates from the model incorporating data about the number of patients as a predictor obtained a median RMSE of 0.67 (minimum, 0.53; maximum, 0.87; n = 10 towns) and a median correlation of 0.86 (minimum, 0.77; maximum, 0.93; n = 10 towns). Figures S2 and S3 show forecasts from both models across all 10 towns.

Table 1 shows importance values for each of the predictors used in the 2-week forecast ensemble model. Because fivefold cross-validation was performed, each value is an average of the five values respective to each of the five models. Calculating variable importance in an ensemble model is difficult because the importance of a variable may be due to its interaction with other variables. In the random forest algorithm, IncNodePurity for a variable is the total decrease in node impurities from splitting on the variable, averaged over all trees. This is done by measuring the residual sum of squares. The first column in Table 1 contains the importance values of predictors when the complete data set was used in designing the ensemble model. The second column in Table 1 contains the importance values of predictors when the data of July to November (the monsoon season) were used. Both models suggest that call volume is the most important predictor of the suspected dengue cases. The relative importance of call volume to weather parameters is higher during the dengue activity season. The relative importance of awareness campaign activities increases in dengue activity season, suggesting that both the number of awareness campaigns starts to increase during the dengue season when cases start to appear. The effect of awareness campaigns on other variables, specifically calling patterns of people, also becomes more significant during the dengue season.

We compare the predictive power of the random forest model with a generalized linear model at both a city- and a subcity-level granularity. For both these models, we used the same training methodology as the random forest (see Materials and Methods) and considered call volume data at the corresponding granularity (city or subcity level). We note that reliable weather parameters are only available at the city level. At the city-level granularity, table S1 compares the individual predictive power of weather parameters and call volume data in estimating the number of suspected dengue cases 2 weeks in the future (on a log scale). The RMSE results suggest that the random forest model significantly outperformed the generalized linear model in predicting future dengue cases. Moreover, the best estimations for the future dengue cases were made by the models that incorporated all the variables as opposed to those that used only a subset of the variables. In addition, the models trained using the combination of calls and awareness level data provide a better estimation of future dengue cases than those trained using weather parameter data.

At a subcity granularity (town), we fit log-suspected cases of 2 weeks in the future using a generalized linear model. The estimates generated from this model obtained a median RMSE of 1.11 (minimum, 0.96; maximum, 1.53; n = 10 towns) and a median correlation of 0.55 (minimum, 0.30; maximum, 0.71; n = 10 towns), with log-suspected cases reported from the government hospitals as shown in fig. S4. In comparison, the random forest model yields significantly lower RMSE values and significantly higher correlation than the linear models.

Finally, we compare the predictive power of a subcity-level model based entirely on weather parameter data with our original model. Log-suspected cases of 2 weeks in the future for each town were fit separately using a random forest model. The estimates generated from the model obtained a median RMSE of 1.14 (minimum, 0.84; maximum, 1.35; n = 10 towns) and a median correlation of 0.51 (minimum, 0.26; maximum, 0.65; n = 10 towns), with log-suspected cases reported from the government hospitals as shown in fig. S5. These results show that weather parameter data alone do not have enough predictive power to provide good dengue forecasts.

Deployment of a live system

Our phone-based dengue surveillance system described in this paper has been successfully deployed in collaboration with the government of Punjab in Pakistan, and the results are widely disseminated through the disease activity dashboard used by public health officials. A web service API call retrieves data from the backend database and performs statistical analyses offline through batch mode on a weekly basis to make predictions of 2 weeks in the future. To incorporate the changes in behavior of citizens calling in, each week with the addition of newer data, the model is retrained to incorporate the changes in behavior of users. The predictions are then generated on the basis of the newly trained model.

Figure 4 shows the different snapshots of the health hotline interface used by the operators and public health officials. The activity dashboard helps them visualize quickly and evaluate the towns that are most vulnerable to an increase in suspected cases and thereby effectively allocate field workers who perform targeted containment activity. Field workers in Punjab are equipped with smartphones by the government. A smartphone application developed by the government of Punjab allows field workers to geo-tag the location and type of containment activity they have performed. This helps government officials ensure that the allocated task has indeed been fulfilled by the field worker. It also helps the officials visualize the locations of the activities and allocate additional workforce if required. Moreover, during the weekly dengue meetings, headed by the chairman of the Punjab government, the health officials of the most vulnerable towns are required

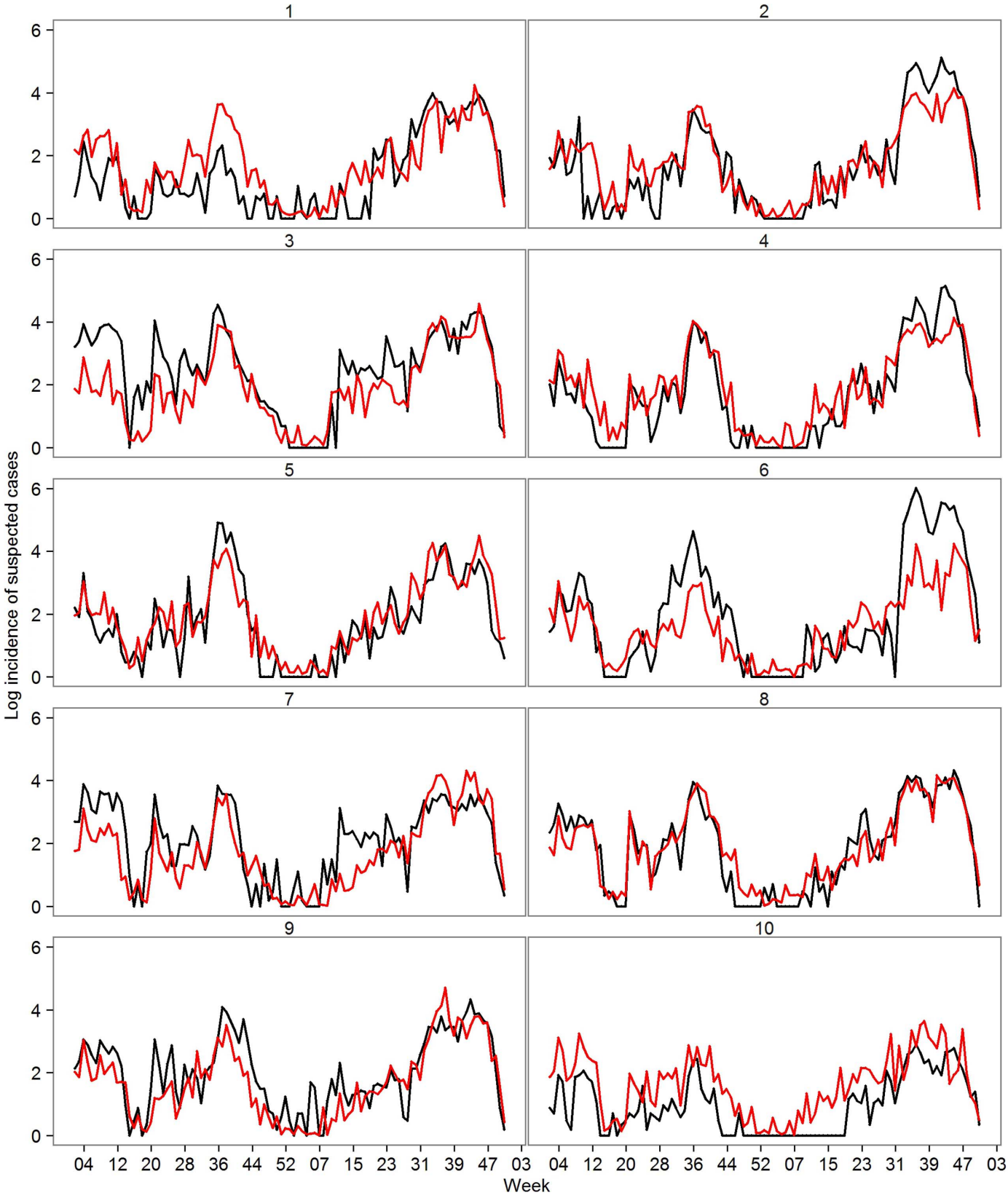

Fig. 2. Town-wise predictions of log-suspected cases from the ensemble model based on calls and weather data. Suspected dengue cases (black) and predictions from the model (red).

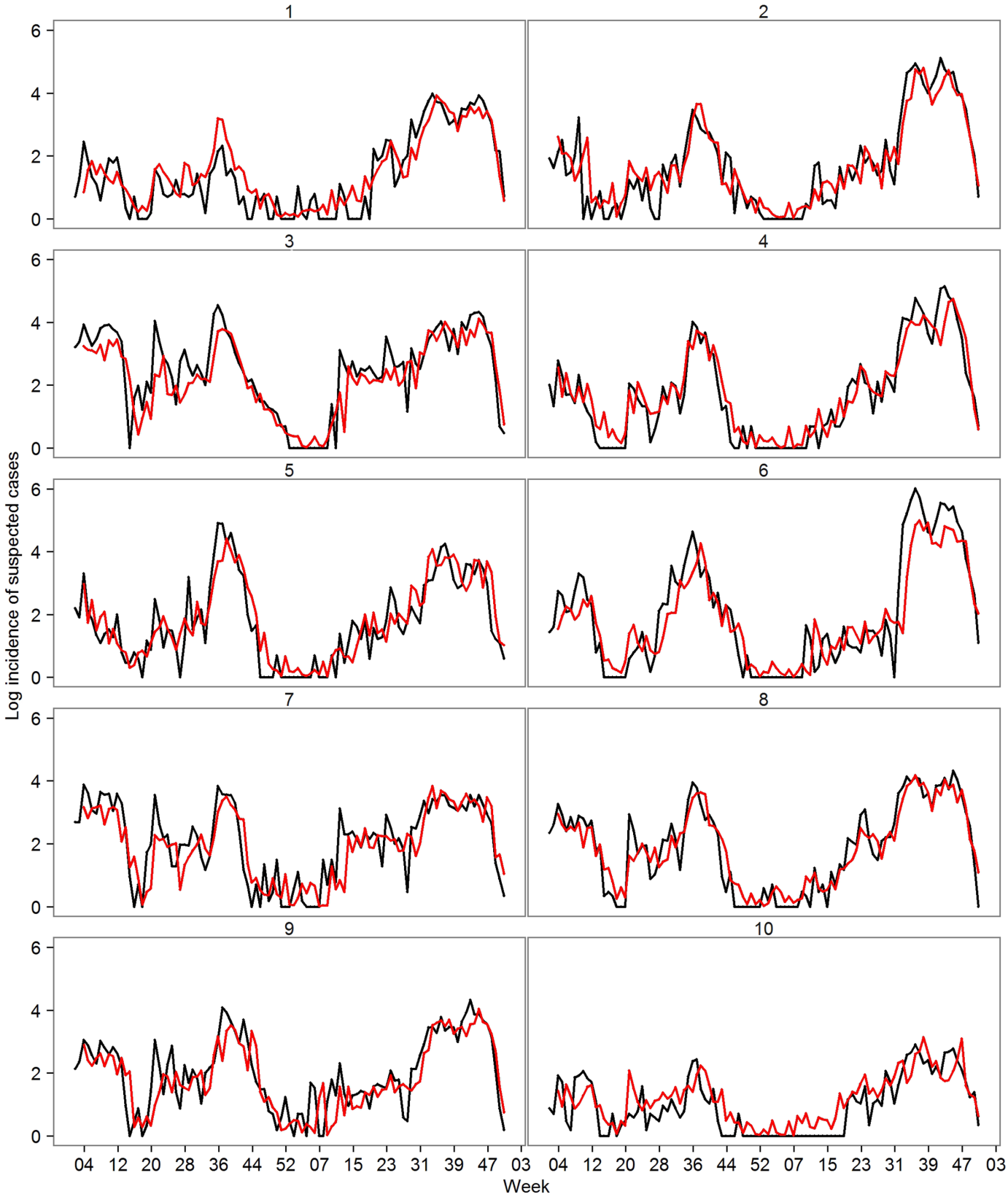

Fig. 3. Town-wise predictions of log-suspected cases from the ensemble model based on calls, cases, and weather data. Suspected dengue cases (black) and predictions from the model (red).

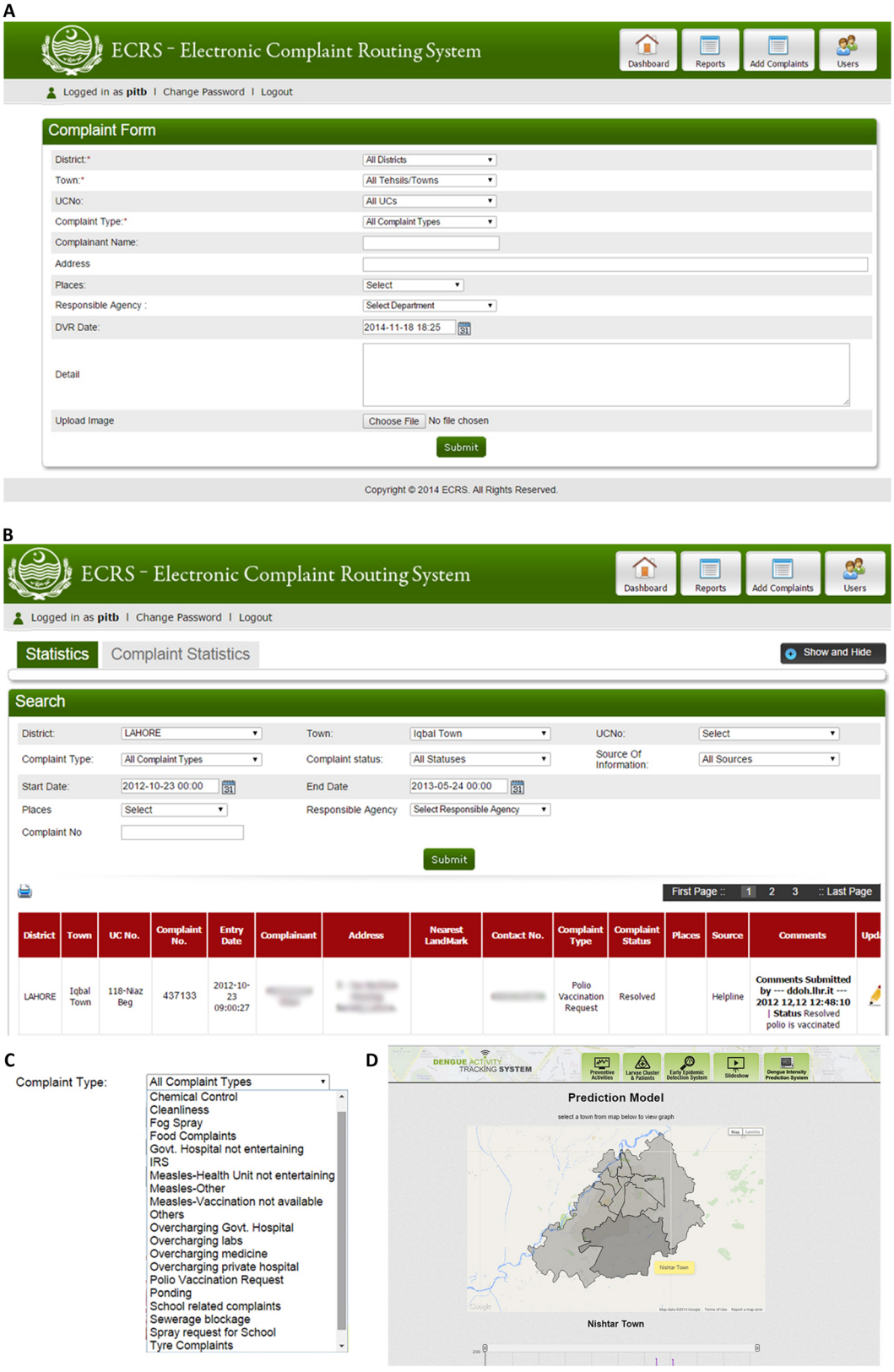

Fig. 4. Punjab Health Hotline Reporting System. (**A**) Interface used by operators to lodge complaints. (**B**) Interface used by officials to view complaints. (**C**) Types of complaints being lodged in the system. (**D**) Front-end interface of our dengue cases prediction system.

to present a detailed report on their efforts to contain the disease in their towns. The predictions from our system are widely disseminated to public health officials and form a critical part of active government strategies for dengue containment. Since the introduction of these containment strategies, we have witnessed a reduction in the number of confirmed dengue patients in Lahore from 21,000 in 2011 to just 257 in 2012 and 1600 in 2013.

DISCUSSION

On the basis of the results, we conclude that call volume data from a simple hotline facility combined with widely available city-level weather data can serve as a good predictor of future suspected dengue cases at a fine-grained subcity level. Surprisingly, adding data about the number of patients into the existing model has a marginal effect in improving the prediction accuracy. The addition of data about the number of patients only provides improvement in capturing the true peak of dengue activity during the high dengue activity season. The appeal of our model is its usefulness despite its sheer simplicity; a simple phone-based health hotline can be used to forecast the number of patients at a subcity location granularity 2 to 3 weeks ahead of time.

Setting up a health hotline to support disease surveillance mechanisms has several advantages. The health hotline facility currently operational is cost-effective, making it ideal for resource-constrained environments in developing countries, such as Pakistan. Telephone triage–based surveillance systems also allow governments to identify disease activity at subcity granularities, leading to effective utilization of their limited health field workers for targeted containment. The forecasts generated from such a system can become a substitute to the paper-based patient report data in outbreak detection systems, which may take weeks to compile. The health hotline can be easily extended to monitor multiple diseases without any substantial increase in allocation of resources; the current health hotline deployed in Pakistan is also being used to monitor polio disease cases. Finally, health hotlines can provide an easy and centralized interface to gather patient reports from hospitals.

Despite the extensive utility of the calls to a health hotline, systems based on call volume have to address several challenges carefully. First, the calls to the health hotline included in our analysis are not exclusively made by dengue patients; patients suffering from diseases with similar symptoms, such as malaria, can contribute to an increase in call volume. Therefore, it is important to devise criteria carefully for identifying calls pertaining to a target disease. Moreover, the use of a health hotline largely depends on the amount of awareness present in a community about the health hotline. Hence, regular awareness campaigns are essential to promote the continuous use of the health hotline. These awareness campaigns are often in the form of radio and television (TV) commercials, banner postings, and awareness group meetings. At the same time, as in our model, the level of awareness needs to be incorporated separately in the disease forecasting models. Given the fact that patients who suffer from a particular strain of dengue virus become immune to it, we assume that most of the callers to the health hotline are first-time callers. The same cannot be said about diseases that are reoccurring in the same individuals. Hence, further studies need to be carried out to find the exact relation between the increase in awareness level in an individual or a community and the decrease in the number of calls to the health hotline. Nevertheless, we suggest continuous retraining of the model with the latest data sets to account for changes in unknown variables over time. Last, calling patterns within a city for individual localities may vary because of the difference in socioeconomic conditions of the inhabitants. Given that the socioeconomic data of inhabitants are generally not available in developing countries, we recommend training separate models for each locality.

In summary, our work demonstrates how a resource-constrained developing country, such as Pakistan, can effectively use a health hotline–based system to provide accurate dengue case forecasts, at fine-grained granularities, 2 to 3 weeks ahead of time. On the basis of our deployment experiences over a 3-year time period, we have observed the direct impact of the hotline system. Our system has helped public health officials to take early actions to contain the spread of the disease and provide hospitals an early warning of dengue cases in their vicinity. We believe that this system can also be used for a broad array of diseases beyond dengue and can easily be replicated in other developing countries at low costs.

MATERIALS AND METHODS

Health hotline

In response to the 2011 outbreak in Lahore, the provincial government in Punjab, Pakistan launched a multipronged effort to combat the epidemic and improve its surveillance and rapid response system. A phone-based disease helpline system was introduced, and a toll-free number was widely publicized through TV and radio advertisements. There are up to 100 operators during an outbreak period dedicated to serving up to 5000 calls a day. Since the inception of the system in September 2011, more than 300,000 calls have been fielded by the system. Callers use the health hotline to inquire if the symptoms they are having are the symptoms for dengue disease. If this is the case, the first responder guides them to the nearest hospitals and provides a brief description of the tests to be conducted and bed availability in the hospitals. In addition, callers can request insecticide spray at homes or in neighborhoods. The helpline has also been used to report sewage leaks, stagnant water accumulation, and overcharging for hospital treatment. The first responders assign a category to each call. These requests are used by public health workers to prioritize sanitation and garbage collection drives. The health hotline is accessible throughout the year to serve as an information portal and to collect real-time citizen feedback. The operators of the health hotline are trained by medical personnel with detailed instructions to handle calls and manuals to identify symptoms of dengue. During the dengue season, doctors are also available at the health hotline, and the operators are instructed to forward the calls to the doctors in case of any ambiguity.

Awareness campaigns

As part of the initiative, the health workers in the government of Punjab were tasked to visit neighborhoods and spread awareness about dengue fever. Awareness seminars are carried out in mosques, schools, and other community settings to inform the general public about the symptoms, spread, and prevention of dengue. The health hotline number is widely disseminated during these seminars as a means to discuss symptoms or to inquire about bed availability in hospitals, if someone is suspected of having dengue. Although these seminars are carried out throughout the year, their frequency increases during the high dengue activity season. Records for each awareness seminar are kept by the government for future allocation of workers to spread awareness in a given town.

Hospitals and definition of cases

During the dengue outbreak, the government institutionalized a regime where any patient suspected of dengue is sent to a public sector hospital for further tests and treatment. These public sector hospitals from Lahore

admitted dengue patients in specialized dengue wards and are accessible to the general public. The government placed three computer operators in each hospital to enter the data of dengue patients in a centralized patient tracking system. Criteria were devised by the Dengue Expert Advisory Group (DEAG) in Pakistan to identify suspected dengue patients so they could be referred for laboratory tests. According to the criteria, if a patient shows three or more of the following symptoms, the patient will be marked as a "suspected dengue case." These symptoms include fever of 2 to 10 days, retro-orbital pain, myalgia, arthralgia/severe backache, rash, bleeding manifestations (epistaxis, hematemesis, bloody stools, menorrhagia, and hemoptysis), abdominal pain, decreased urinary output despite adequate fluid intake, and irritability in infants.

Methods

For our analysis, we used data from the Punjab government's toll-free health hotline service. A single call unit, included in our analysis, is defined as the call received at the health hotline, categorized by the operator as a dengue symptoms inquiry call. The location of the caller is based on the information either provided by the caller or inferred by the first responder using the address given by the caller. The operators are also allowed to mark a call as irrelevant; these calls were excluded from our analysis. Dengue awareness campaigns data were recorded from the health hotline. A single awareness activity unit is defined as a dengue awareness seminar carried out by a health worker from the government and reported at the health hotline. Weather data used in our analysis were retrieved from the Punjab meteorological department. Unlike the calls and awareness level data, which are available at subcity location granularity, only coarse-grained city-level weather parameter data were available. Finally, the number of suspected dengue case data was retrieved from all public sector hospitals in Lahore between 2012 and 2013. A suspected dengue patient is defined as a patient who shows symptoms of dengue and is referred for a standardized laboratory test by the doctor. The location of the patient is determined on the basis of the residence address provided by the patient at the hospital.

Our dengue forecasting system is based on an ensemble model. We use call volume data divided across 10 towns in the city of Lahore. Widely available city-level weather parameters, specifically humidity, rainfall, and temperature, are used in the model to incorporate the seasonality of diseases such as dengue (*21*). Finally, because health hotline awareness among citizens can affect calling patterns, our system treats awareness as a separate parameter using the number of recorded hotline publicity activities in a given time period as proxy. It is important to highlight that previous systems, such as Google Flu Trends and NHS Direct, that purely use query volume data for prediction, may generate inaccurate estimates of an outbreak because they fail to incorporate critical awareness level data in their model (*18*). Awareness level can be defined as the likelihood that a person will use an alternative source, such as a search engine or a health hotline, to inquire about symptoms of a disease at a given time of a year. Specific to search queries, this can vary for several reasons; if a new disease-related drug is introduced in the market or if a celebrity gets infected with a disease, people are more likely to search online about the disease, contributing to an increase in searches. This leads to an overestimation of disease activity, which is a common problem in previous systems (*18*). Hence, awareness level data need to be incorporated in the model to cater for varying population interest in the alternative data source. Moreover, it is also important to note that although an increase in search queries and internet awareness can be a result of a wide range of factors, increasing awareness of a health hotline is largely dependent on the publicity it receives during the awareness campaigns. Because these awareness campaigns and their locations can be easily monitored, they can be easily accounted for in our model. Finally, the suspected dengue case data were taken from the hospitals in Punjab to serve as the response variable in our model.

Our ensemble model uses the random forest learning algorithm. The choice to use a random forest–based learning algorithm was largely guided by the advantages over linear regression/classification methods. Given the nonlinearity observed in the data, the random forest algorithm significantly mitigates the possibility of overfitting by using an ensemble collection of decision trees and randomizing over features used in training each individual decision tree (see Fig. 1 and fig. S4). Random forests also allow for easy interpretability of the variable importance of features. Although other more sophisticated models, such as gradient-boosted decision trees, were viable alternatives, the performance gains we achieved using random forests combined with the convenience of training and deployment into our online outbreak detection system led us to favor random forests.

Between 1 January 2012 and 31 December 2013, data from the city of Lahore were recorded from the mentioned sources. Weekly counts for each town were computed separately. Weekly aggregates of 2 years for 10 towns generated a total of 1030 points. To validate our hypothesis, we performed, fivefold cross-validation. The data points were split into five randomly selected nonoverlapping folds, each containing 206 points. A region-independent random forest model of regression trees comprising 500 trees and three-node sampling was trained using four folds and validated on the held-out fold. The process was repeated until each fold had been validated. The performance of the models was evaluated using RMSE and correlation values. This was done to capture the efficiency of the model in forecasting the exact values and capturing variations in peaks and lows of the predicted values.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/2/7/e1501215/DC1

fig. S1. Cross-correlation between suspected incidences reported at hospitals and calls received at the health hotline in Lahore for the year 2012.

fig. S2. Town-wise predictions of 3-week log-suspected incidence forecast from ensemble model based on calls and weather data.

fig. S3. Town-wise predictions of 3-week log-suspected incidence forecast from ensemble model based on calls, cases, and weather data.

fig. S4. Town-wise predictions of 2-week log-suspected incidence forecast from generalized linear model based on calls and weather data.

fig. S5. Town-wise predictions of 2-week log-suspected incidence forecast from location-dependent ensemble model based on weather data.

table S1. RMSE values between predicted and actual number of log-suspected cases for various models trained on coarse-grained city-level data set.

REFERENCES AND NOTES

1. Population Profile Punjab | Population Welfare Department (Population Profile Punjab | Population Welfare Department), www.pwd.punjab.gov.pk/population_profile [accessed September 1, 2015].
2. M. A. Khan, E. M. Ellis, H. A. Tissera, M. Y. Alvi, F. F. Rahman, F. Masud, A. Chow, S. Howe, V. Dhanasekaran, B. R. Ellis, D. J. Gubler, Emergence and diversification of dengue 2 cosmopolitan genotype in Pakistan, 2011. *PLOS One* **8**, e56391 (2013).
3. C. Ash, Dangerous dengue provocation. *Sci. Signal.* **3**, ec145 (2010).
4. National Institute of Allergy and Infectious Diseases, Dengue fever treatment; www.niaid.nih.gov/topics/denguefever/understanding/pages/treatment.aspx [accessed September 1, 2015].
5. World Health Organization, Planning social mobilization and communication for dengue fever prevention and control; www.who.int/tdr/publications/documents/planning_dengue.pdf [accessed September 1, 2015].
6. Centers for Disease Control and Prevention, Dengue symptoms and treatment; www.cdc.gov/dengue/symptoms/ [accessed September 1, 2015].

7. N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, D. S. Burke, Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437**, 209–214 (2005).
8. I. M. Longini Jr., A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. T. Cummings, M. E. Halloran, Containing pandemic influenza at the source. *Science* **309**, 1083–1087 (2005).
9. V. Racloz, R. Ramsey, S. Tong, W. Hu, Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. *PLOS Negl. Trop. Dis.* **6**, e1648 (2012).
10. W. K. Yih, K. S. Teates, A. Abrams, K. Kleinman, M. Kulldorff, R. Pinner, R. Harmon, S. Wang, R. Platt, Telephone triage service data for detection of influenza-like illness. *PLOS One* **4**, e5260 (2009).
11. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
12. E. H. Chan, V. Sahai, C. Conrad, J. S. Brownstein, Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLOS Negl. Trop. Dis.* **5**, e1206 (2011).
13. B. M. Althouse, Y. Y. Ng, D. A. T. Cummings, Prediction of dengue incidence using search query surveillance. *PLOS Negl. Trop. Dis.* **5**, e1258 (2011).
14. C. C. Freifeld, K. D. Mandl, B. Y. Reis, J. S. Brownstein, HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J. Am. Med. Inform. Assoc.* **15**, 150–157 (2008).
15. S. Hales, N. de Wet, J. Maindonald, A. Woodward, Potential effect of population and climate changes on global distribution of dengue fever: An empirical model. *Lancet* **360**, 830–834 (2002).
16. A. Doroshenko, D. Cooper, G. Smith, E. Gerard, F. Chinemana, N. Verlander, A. Nicoll, Evaluation of syndromic surveillance based on National Health Service Direct derived data–England and Wales. *Morb. Mortal. Wkly. Rep.* **54**, 117–122 (2005).
17. D. L. Cooper, G. Smith, M. Baker, F. Chinemana, N. Verlander, E. Gerard, V. Hollyoak, R. Griffiths, National symptom surveillance using calls to a telephone health advice service—United Kingdom, December 2001–February 2003. *Morb. Mortal. Wkly. Rep.* **53**, 179–183 (2004).
18. D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLOS Comput. Biol.* **9**, e1003256 (2013).
19. S. Runge-Ranzinger, P. J. McCall, A. Kroeger, O. Horstick, Dengue disease surveillance: An updated systematic literature review. *Trop. Med. Int. Health* **19**, 1116–1160 (2014).
20. B. M. Althouse, S. V. Scarpino, L. A. Meyers, J. W. Ayers, M. Bargsten, J. Baumbach, J. S. Brownstein, L. Castro, H. Clapham, D. A. T. Cummings, S. Del Valle, S. Eubank, G. Fairchild, L. Finelli, N. Generous, D. George, D. R. Harper, L. Hébert-Dufresne, M. A. Johansson, K. Konty, M. Lipsitch, G. Milinovich, J. D. Miller, E. O. Nsoesie, D. R. Olson, M. Paul, P. M. Polgreen, R. Priedhorsky, J. M. Read, I. Rodríguez-Barraquer, D. J. Smith, C. Stefansen, D. L. Swerdlow, D. Thompson, A. Vespignani, A. Wesolowski, Enhancing disease surveillance with novel data streams: Challenges and opportunities. *EPJ Data Sci.* **4**, 17 (2015).
21. Centers for Disease Control and Prevention, www.cdc.gov/Dengue/entomologyEcology/climate.html [accessed September 1, 2015].

Acknowledgments: We thank the government of Punjab and the Punjab Information Technology Board for initiating this project and providing funding and logistical support for this project. We thank the Pakistan Meteorological Department for sharing the weather data sets. **Funding:** N.A.R. and T.A. were supported as research assistants by Information Technology University, Pakistan, and New York University (NYU) during the course of this project. We thank the NYU Abu Dhabi Research Institute and the Center for Technology and Economic Development (CTED) at NYU Abu Dhabi for providing funding support for L.S., T.A., and S.K. in this project. S.K. was also partially supported as a postdoctoral fellow by an NSF grant. F.P. is supported on a doctoral fellowship by the University of Washington. **Author contributions:** T.A., F.P., N.A.R., and U.S. worked closely with the Punjab Information Technology Board to deploy the system, collect the data, and disseminate the results. N.A.R., S.K., L.S., and U.S. designed the study, methods, data analysis, and results. N.A.R., L.S., U.S., and S.K. wrote the manuscript. L.S. and U.S. guided the project and edited the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data, at an aggregated level, needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Environmental data is publicly available in Pakistan Meteorological department's website. Call volume data, awareness campaign data, and dengue cases data are the property of the government of Punjab, Pakistan. Under the agreement with the government of Punjab, these data sets may be requested from the authors.

Submitted 1 September 2015
Accepted 15 June 2016
Published 8 July 2016
10.1126/sciadv.1501215

Citation: N. Abdur Rehman, S. Kalyanaraman, T. Ahmad, F. Pervaiz, U. Saif, L. Subramanian, Fine-grained dengue forecasting using telephone triage services. *Sci. Adv.* **2**, e1501215 (2016).

ScienceAdvances
AAAS

Fine-grained dengue forecasting using telephone triage services

Nabeel Abdur Rehman, Shankar Kalyanaraman, Talal Ahmad, Fahad Pervaiz, Umar Saif and Lakshminarayanan Subramanian

(July 8, 2016)
Sci Adv 2016, 2:.
doi: 10.1126/sciadv.1501215