

# Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore

Yuan Shi,<sup>1</sup> Xu Liu,<sup>1</sup> Suet-Yheng Kok,<sup>1</sup> Jayanthi Rajarethinam,<sup>1</sup> Shaohong Liang,<sup>1</sup> Grace Yap,<sup>1</sup> Chee-Seng Chong,<sup>1</sup> Kim-Sung Lee,<sup>1</sup> Sharon S.Y. Tan,<sup>2</sup> Christopher Kuan Yew Chin,<sup>1</sup> Andrew Lo,<sup>3</sup> Waiming Kong,<sup>4</sup> Lee Ching Ng,<sup>1,5</sup> and Alex R. Cook<sup>6,7</sup>

<sup>1</sup>Environmental Health Institute, <sup>2</sup>Environmental Public Health Operations Department, and <sup>3</sup>Centre for Climate Research Singapore, National Environment Agency, Singapore; <sup>4</sup>School of Engineering, Nanyang Polytechnic, Singapore; <sup>5</sup>School of Biological Sciences, Nanyang Technological University, Singapore; <sup>6</sup>Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore; <sup>7</sup>Yale-NUS College, National University of Singapore, Singapore

**BACKGROUND:** With its tropical rainforest climate, rapid urbanization, and changing demography and ecology, Singapore experiences endemic dengue; the last large outbreak in 2013 culminated in 22,170 cases. In the absence of a vaccine on the market, vector control is the key approach for prevention.

**OBJECTIVES:** We sought to forecast the evolution of dengue epidemics in Singapore to provide early warning of outbreaks and to facilitate the public health response to moderate an impending outbreak.

**METHODS:** We developed a set of statistical models using least absolute shrinkage and selection operator (LASSO) methods to forecast the weekly incidence of dengue notifications over a 3-month time horizon. This forecasting tool used a variety of data streams and was updated weekly, including recent case data, meteorological data, vector surveillance data, and population-based national statistics. The forecasting methodology was compared with alternative approaches that have been proposed to model dengue case data (seasonal autoregressive integrated moving average and step-down linear regression) by fielding them on the 2013 dengue epidemic, the largest on record in Singapore.

**RESULTS:** Operationally useful forecasts were obtained at a 3-month lag using the LASSO-derived models. Based on the mean average percentage error, the LASSO approach provided more accurate forecasts than the other methods we assessed. We demonstrate its utility in Singapore's dengue control program by providing a forecast of the 2013 outbreak for advance preparation of outbreak response.

**CONCLUSIONS:** Statistical models built using machine learning methods such as LASSO have the potential to markedly improve forecasting techniques for recurrent infectious disease outbreaks such as dengue.

**CITATION:** Shi Y, Liu X, Kok SY, Rajarethinam J, Liang S, Yap G, Chong CS, Lee KS, Tan SS, Chin CK, Lo A, Kong W, Ng LC, Cook AR. 2016. Three-month real-time dengue forecast models: an early warning system for outbreak alerts and policy decision support in Singapore. *Environ Health Perspect* 124:1369–1375; <http://dx.doi.org/10.1289/ehp.1509981>

## Introduction

Dengue is an acute infectious disease common to tropical and subtropical regions. Dengue viruses are transmitted by *Aedes* mosquitoes, mainly *Aedes aegypti* and *Aedes albopictus* (Rosen et al. 1983). Globally, the World Health Organization has estimated that there are 50–100 million dengue infections per year (Rigau-Pérez et al. 1998), although more recent estimates have elevated this figure to 390 million, of which ~96 million are symptomatic (Bhatt et al. 2013). Dengue infection in humans is mostly self-limiting—although antiviral drugs are under development (Lim et al. 2013; Rathore et al. 2011)—but may require hospital admission, and the more severe manifestations of dengue may lead to death (Murphy and Whitehead 2011). Case fatality rates of dengue fever and severe dengue vary from 0–5% to 3–5% (Halstead 1999).

The city-state of Singapore, which lies approximately 130 km north of the equator, has a tropical rainforest climate in the

Köppen–Geiger climate classification system (Peel et al. 2007) with no distinctive seasons. The climate, combined with Singapore's highly urbanized environment, favors the presence of *Aedes* mosquitoes and the transmission of dengue virus (Thu et al. 1998), thus making Singapore highly vulnerable to dengue outbreaks. All four serotypes are endemic to Singapore, and there is frequent introduction and circulation of different genotypes of the virus (Lee et al. 2010, 2012). With an annual reported incidence in the range of 20–330 per 100,000 people, the economic impact of dengue in Singapore from 2000 to 2010 was estimated to be 850 million USD–1.15 billion USD, or approximately 200 USD per capita per year (Carrasco et al. 2011). Since 2003, > 100 dengue-related deaths have been reported [Ministry of Health, Singapore (MOH) 2012].

Antiviral drugs and vaccines have yet to reach the market (Douglas et al. 2013), and initial results from trials have been discouraging

(Halstead 2012; Mahalingam et al. 2013; Sabchareon et al. 2012). In the absence of an effective vaccine against dengue, suppressing the mosquito vector population remains the key thrust of Singapore's dengue-control program (Lee et al. 2013). From 2000 to 2009, the country spent > US\$85 million per annum (Carrasco et al. 2011) on this endeavor. Since 2006, the National Environment Agency (NEA) has introduced virological surveillance for early warning of outbreaks (Lee et al. 2010) and a novel mosquito-breeding index that estimates the spatial distribution of *Ae. aegypti*, the main dengue vector in Singapore. Previous predictive capability relied on a qualitative understanding based on temperature, circulating serotype, vector data from the field, and estimated immunity level of the human population. If automated, statistical models hold the promise of being able to provide real-time quantitative forecasts of the appearance and evolution of a dengue outbreak, which may be used to efficiently guide the deployment of vector-control operations.

Address correspondence to L.C. Ng, Environmental Health Institute, National Environment Agency, 11 Biopolis Way #06-05/08 Helios Block, Singapore, S138667. Telephone: (65) 6771 9108. E-mail: NG\_Lee\_Ching@nea.gov.sg

Supplemental Material is available online (<http://dx.doi.org/10.1289/ehp.1509981>).

We thank H.-K. Tan for leading the research at its initial stage. We also thank other colleagues at the National Environment Agency, Singapore; Communicable Disease Division of the Ministry of Health, Singapore; and the Department of Statistics, Singapore, for valuable comments, information, and support.

This work is supported by the Environmental Initiative Fund from the Ministry of the Environment and Water Resources, Singapore. A.R.C. also received funding support from the Centre for Infectious Disease Epidemiology and Research, Singapore; the Ministry of Defence, Singapore; the Ministry of Education, Singapore; the Ministry of Health, Singapore, from the Health Services Research Grant (HSRG12MAY023) and the Communicable Disease Public Health Research Grant (CDPHRG12NOV021); and from the National University Health System, Singapore.

The authors declare they have no actual or potential competing financial interests.

Received: 20 March 2015; Accepted: 24 November 2015; Advance Publication: 11 December 2015; Final Publication: 1 September 2016.

Any statistical approach to forecast dengue would need to meet certain criteria to be practical: *a*) use only data that are available at the time the forecast is made; *b*) be capable of forecasting weeks or months into the future to give lead time for preparing a public health response (for instance, hiring new control staff); *c*) possess validated and demonstrated predictive performance using data that were not used in its construction, to prevent over-fitting and to ascertain confidence levels; and lastly, *d*) be able to process new data rapidly. Population dynamic modeling of dengue exploits epidemiological and entomological knowledge and is valuable for modeling what-if scenarios, such as the effect of introducing changes to the host or mosquito population [such as introducing vaccination into the pediatric vaccination schedule (Coudeville and Garnett 2012) or releasing genetically modified sterile mosquitoes or those infected by *Wolbachia* (Hughes and Britton 2013)], and may be useful for predicting long-term changes to epidemic dynamics caused by changing levels of herd immunity or by the age structure of a population (Cummings et al. 2009; Egger et al. 2008). A limitation of population dynamic models is that they are difficult to integrate with real-time data streams, such as meteorological or incidence data. Although success has been achieved for epidemiologically simpler diseases such as influenza (Baguelin et al. 2010; Ong et al. 2010), the complexity caused by the possibility of having several serotypes circulating simultaneously and by the influence of the environment on the vector makes these models a relatively unattractive choice for short-term forecasts of dengue. In contrast, correlative statistical approaches—which describe the phenomenon but not the underlying process—are well suited to integration with multiple live data streams and may have good predictive accuracy if future conditions do not stray too far from the conditions used to parameterize them.

Other researchers have sought to develop statistical time series models to predict dengue in Singapore. Earnest et al. (2012) compared 1-week-ahead dengue forecasts based on two popular modeling methods—the autoregressive integrated moving average (ARIMA) and the Knorr-Held two-component (K-H) model—and showed that the K-H model was slightly more accurate. Hii et al. (2012a, 2012b) developed a forecasting model using Poisson multivariate regression to predict the number of cases over a 4-month interval, demonstrating that past temperature and rainfall data are good predictors of future dengue incidence. Another study (Althouse et al. 2011) predicted 1-week and 1-month dengue incidence in Singapore and Bangkok, respectively, based on internet search engine

queries that might signal the early stages of an outbreak. The authors compared three statistical approaches based on these data—step-down linear regression, generalized boosted regression, and negative binomial regression—and concluded that the step-down linear model was superior to the other two. Although the aforementioned models meet many of the criteria noted above, it is noteworthy that none has been validated against data not used in its construction, and none was developed explicitly for operational use, suggesting that their predictive performance and usefulness to operations were not tested.

In this paper, we describe a new approach to forecasting dengue that is used by Singapore's NEA in planning vector control and in public communication. The model specifically optimizes predictive accuracy over a 3-month time horizon with model complexity selected, and predictive performance evaluated, using out-of-sample forecasting. We show that this approach, which uses the least absolute shrinkage and selection operator (LASSO) method to fit large regression models, has better predictive performance than other modeling approaches.

## Materials and Methods

### Statistical Analyses

**LASSO.** The Least Absolute Shrinkage and Selection Operator (LASSO) is a technique that was proposed in the 1990s (Tibshirani 1996) and has inspired much interest in the statistical methodology community on “small  $n$  large  $p$ ” problems (Tibshirani 1996). This framework extends standard regression and related models such as logistic regression by simultaneously selecting which parameters to include in the model and what their values should be. Rather than optimizing the (log) likelihood  $L(y|\beta, x)$  for dependent variable  $y$ , independent variables  $x$  and coefficients  $\beta$ , as in standard regression, LASSO optimizes the sum of the log-likelihood and a penalty term controlled by an additional parameter  $\lambda$ , which controls model complexity. In particular, the optimal coefficients are the  $\beta$ s that maximize

$$L(y|\beta, x) - \lambda \sum_k |\beta_k|. \quad [1]$$

The penalty term  $\lambda$  controls model complexity: for a specific value of the penalty parameter, the optimal fit will have some (or many) coefficients set to 0, that is, some covariates are not used in the model. The penalty term is usually selected by cross-validation, in which *a*) the data are partitioned into several training and test sets; *b*) for each training set, a series of values of  $\lambda$  are considered; and *c*) for each  $\lambda$ , the best values of  $\beta$  are found and then used to predict

the test data so that the out-of-sample performance can be measured. The value of  $\lambda$  that maximizes the average out-of-sample performance is then used to select the final model, implicitly meaning that the model complexity is selected to optimize predictive accuracy for a given set of independent variables. We used the glmnet algorithm (Friedman et al. 2010) implemented in the R statistical language (R Core Team 2014) to fit the models.

Our approach was to develop a tailored submodel unique to each forecast window from 1 week to 12 weeks in the future. For each submodel, the outcome variable was the weekly number of cases (natural log-transformed, with 1 added to avoid logging 0), and a large set of potential input variables were considered (details may be found in the data section). The formulation for each submodel was a multivariable linear regression. LASSO was used to select a (potentially different) set of predictors for each forecast window along with the values of their coefficients, with 10-fold cross-validation used to determine optimal model complexity, before the forecasts were “stitched together” graphically to create the impression of a single predictive routine.

Covariates were considered at lags of up to 20 weeks based on the findings of a previous study by Hii et al. (2012a, 2012b), but in contrast to their approach, we allowed the effect of a single factor (such as temperature) to have multiple lags in influencing future dengue cases. The framework used in developing the models is presented in Figure 1.

### Other Approaches

We applied statistical methods used by Earnest et al. (2012) and Althouse et al. (2011) to compare with the LASSO approach described above.

In the Seasonal ARIMA (SARIMA) algorithm, models are composed of nonseasonal factors ( $p, d, q$ ) and seasonal factors ( $P, D, Q$ ), where  $d$  and  $D$  define the order (i.e., the number of weeks in the past) of nonseasonal and seasonal differencing in the time series (between successive values, used to reduce the effects of nonstationarity of the time series),  $p$  and  $P$  are the autoregressive (AR) terms, and  $q$  and  $Q$  are the moving average terms. SARIMA models can vary from very simple—for instance, a nonseasonal AR1 model in which the dependent variable is regressed upon itself ( $y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$ )—to very complex, where the dependent variable depends on several past weeks, on moving averages of past weeks' data, and on recurrent seasonal factors, in the present case, 52 weeks ago. The values of ( $p, d, q, P, D$ , and  $Q$ ) with the lowest Akaike information criterion (AIC) during model training are selected for the optimal model. [A definitive introduction is provided by Chatfield (2013).]

In the step-down linear regression (LR) model, we developed a submodel for each forecast window as in the LASSO approach, starting with a model containing all predictors (at multiple lags) and progressively eliminating variables one at a time according to AIC score, until no further improvement was possible without removing two or more terms simultaneously.

### Model Comparison

Models were compared using the mean absolute percentage error (MAPE), as proposed by Armstrong (1985) and modified and reposed by Flores (1986), for each forecast window. If equation  $D_{t+w}$  is the actual number of dengue cases  $w$  weeks after time  $t$  when the prediction is made, and  $D_{t+w}^m$  is the number of cases forecasted by model  $m$ , the MAPE for that model and forecast window is

$$MAPE(m, w) = \frac{100\%}{|V|} \sum_{t \in V} \left( \frac{|D_{t+w} - D_{t+w}^m|}{D_{t+w}} \right), \quad [2]$$

where  $V$  is the validation set. We used data from 2001 to 2010 as training data to parameterize the models and from 2011 to 2012 to validate the models, and the validated models were applied to 2013 data to evaluate their performance. The same training and validation data sets were used for all three models.

Predictions were accompanied by 95% intervals using point estimates from the models with overlaid error using the estimated standard deviation of residuals from the fitted model, and we calculated the coverage of these

intervals, that is, the proportion of time points in which the eventual data point fell within the prediction interval, to compare whether the actual coverage of the three approaches was at or above the nominal coverage.

### Data

Weekly covariates were used to match the frequency of reported dengue data made available by Singapore's Ministry of Health, with finer resolution data coarsened to this level, and coarser data assumed to be homogeneous over each reporting interval. The time horizon used for all variables was January 2001 to December 2012 unless otherwise indicated. The variables used and their sources are described below:

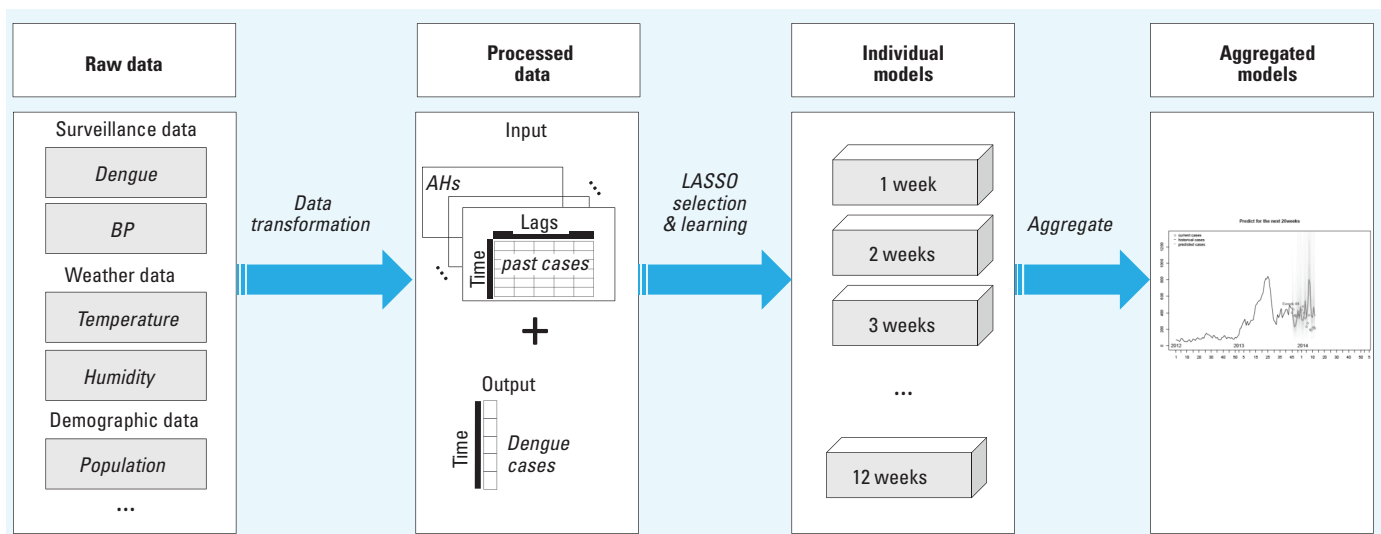
**Case data.** The weekly number of cases (natural log-transformed, +1) was provided by the Ministry of Health, Singapore, and can be obtained from their Weekly Infectious Diseases Bulletin ([https://www.moh.gov.sg/content/moh\\_web/home/statistics/infectiousDiseasesStatistics/weekly\\_infectiousdiseasesbulletin.html](https://www.moh.gov.sg/content/moh_web/home/statistics/infectiousDiseasesStatistics/weekly_infectiousdiseasesbulletin.html)).

**Population data.** Midyear population sizes for residents and foreign non-residents were obtained for each year from the Singapore Department of Statistics (Department of Statistics, Singapore 2015), natural log-transformed, and applied to all weeks within the corresponding calendar year. These data were applied to all weeks within that calendar year.

**Meteorological data.** Weekly mean temperature ( $T$ ) in degrees Celsius, maximum hourly temperature, number of hours of high temperature ( $> 27.8^\circ\text{C}$ ) each week, and

weekly relative humidity (RH) were obtained from Meteorological Services Singapore. Absolute humidity for any week ( $H_A$ ) was calculated from the weekly mean temperature ( $T$ ) and the relative humidity ( $H_R$ ) using standard formulae (Xu et al. 2014).

**Vector surveillance data.** The weekly breeding percentage (BP) is an in-house index developed by NEA that provides an estimate of the proportion of *Ae. aegypti*, the primary vector of dengue in Singapore, compared with all *Aedes* spp. As part of vector control operations, potential breeding sites are sought, samples are taken when breeding is found, and the species is determined in our laboratory. Because this is part of routine vector control and not solely for surveillance, efforts are not temporally or spatially regular, and they tend to favor outbreak periods and areas with transmission, thus biasing estimates upwards for both total incidence of breeding sites and *Ae. aegypti* breeding. To overcome the biases in data collection, we used the proportion of identified *Ae. aegypti* breeding out of all identified breeding sites to quantify the amount of "relevant" breeding. There are two justifications for this assumption: *a)* another *Aedes* species, *Ae. albopictus*, is so widespread that the amount of *Ae. albopictus* breeding found is a good proxy for total effort in identifying breeding sites, and *b)* *Ae. aegypti* is the primary vector for dengue in Singapore; this species is more efficient at transmission and more often found to be infected than *Ae. albopictus*, and the presence of *Ae. aegypti* is necessary for sustained transmission in any neighborhood. The breeding percentage for week  $t$ ,  $BP_t$ , was calculated from the weekly number of *Aedes*



**Figure 1.** Model framework. Raw data from 2001 to 2010 including dengue cases, breeding percentage, temperature, humidity and population statistics are collated from the Ministry of Health (MOH) (Weekly Infectious Diseases Bulletin; [https://www.moh.gov.sg/content/moh\\_web/home/statistics/infectiousDiseasesStatistics/weekly\\_infectiousdiseasesbulletin.html](https://www.moh.gov.sg/content/moh_web/home/statistics/infectiousDiseasesStatistics/weekly_infectiousdiseasesbulletin.html)), National Environment Agency (<http://www.dengue.gov.sg/>), and Department of Statistics, Singapore (2015) before being transformed and divided into sets that each contain predictors and a single output (dengue incidence at a future time point). In the third step, the LASSO method is employed to train and select the optimal models for future dengue forecast. In total, 12 models are generated, one per forecast week. In the final step, the 12 models are aggregated to make predictions over a 12-week window.



mosquito breeding sites recorded during ground inspections carried out by NEA using the following formula:

$$BP_t = \frac{N_{aegypti}(t)}{N_{total}(t)}, \quad [3]$$

where  $N_{aegypti}(t)$  and  $N_{total}(t)$  are the number of breeding sites containing only *Ae. aegypti* or containing either *Ae. aegypti* or *Ae. albopictus*, respectively, in week  $t$ . This index has been found to have a high correlation with dengue cases (unpublished data) and to be negatively correlated in space with chikungunya cases, for which *Ae. albopictus* is the more competent vector (Ng et al. 2009).

**Trend and seasonality data.** In addition to climatic factors, dengue is affected by other factors such as changes to vector control and circulating serotypes. To address the impact of such nonclimatic factors on disease dynamics, we decomposed dengue incidence into terms for trend and for annual seasonality.

To account for changes, both gradual and abrupt, we extracted trends and seasonality from the weekly time series by using the Breaks For Additive Seasonal and Trend (BFAST) algorithm (Verbesselt et al. 2010). Specifically, BFAST decomposes time series into seasonality and trend components through iterative estimation of time series parameters and detection of break points, delimiting time windows in which different seasonal and trend patterns apply. Within each time window, the effect of seasonality is assumed to be sinusoidal, but the characteristics of the sinusoidal functions vary across time windows. Similarly, trend is defined to be piecewise linear between break points. The inferred seasonal and trend terms were extracted from BFAST and were allowed to be used as covariates in the LASSO model.

## Results

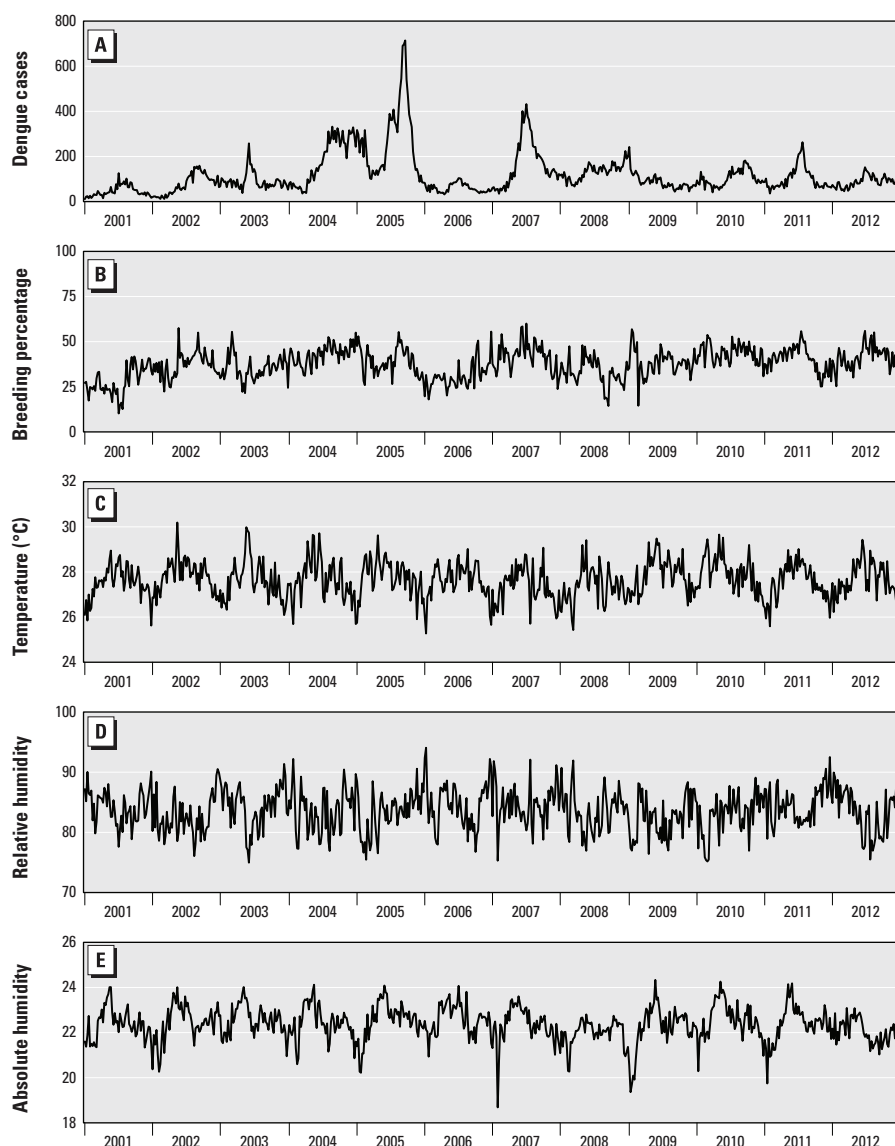
Data are presented in Figure 2. There was little overall variation in weather seasonality in Singapore over the time period investigated (2001 to 2010), with slightly hotter temperatures (around 1–2°C higher) and slightly higher absolute humidity registered in the middle of the year. Relative humidity did not display any notable patterns, and the breeding percentage (i.e., the relative amount of *Ae. aegypti*) varied without any clear pattern. In contrast, dengue fluctuated between low-level endemic and larger epidemic states.

In cross-validation, 12 sets of optimal model complexity parameters were selected for the 12 forecast windows. These included covariates whose effect was lagged from 1 to 20 weeks, and counting each lag separately in a total of 226 data streams, including seasonality and trend. We present the 12-week forecasts, including 95% prediction intervals, for various

time points over the period 2001–2012 in Figure 3 for the LASSO method and the two other methods (step-down linear regression and SARIMA) (Dynamic 12-week forecasts for each model are presented in Video Files S1, S2, and S3). The LASSO and step-down approaches yielded forecasts that more accurately presaged short-term incidence than did the SARIMA model. The start and end of several epidemics were accurately forecasted by both the LASSO and step-down approaches, although the peak of the large 2005 outbreak was not well described by the LASSO model.

The relative forecast accuracy was formally assessed by dividing the dataset into training (2001–2010) and validation sets (2011–2012) and comparing the MAPE of the best guess of the forecast and the coverage of the forecast

interval. The results (Figure 4) support the use of LASSO to construct the forecasts: the LASSO approach yielded more accurate forecast time periods for all except the first 2-week window, in which the performance of the LASSO and step-down approaches were approximately equal. Notably, the MAPE degrades slowly over time under the LASSO approach, with a rise from 17% error [95% confidence interval (CI): 16, 19%] forecasting 1 week to 24% error (95% CI: 22, 26%) forecasting 3 months into the future. In contrast, both the step-down and SARIMA approaches had a MAPE of 29% at 3 months ahead (95% CI: 26, 32% [SARIMA], 27, 32% [step-down]). In addition, although the LASSO and SARIMA predictions were conservative in the sense that the actual



**Figure 2.** Model input from 2001 to 2012. (A) dengue incidence, (B) breeding percentage (BP), defined as the proportion of *Ae. aegypti*, the primary vector of dengue in Singapore, compared with all *Aedes* spp., (C) temperature, (D) relative humidity, and (E) absolute humidity from 2001 to 2012. Data sources are described in the text.

coverage of prediction intervals exceeded the target of 95%, the step-down approach led to forecasts that understated the uncertainty, with a coverage that sometimes fell below the nominal level of 95% (Figure 4).

Interpretation of climatic and other factors was difficult because the strength of their association varied between forecast windows and because they operated over different time lags. However, recent dengue incidence (the autoregressive component) over a lag window of 1–5 weeks generally increased the forecast number of dengue cases; higher average weekly temperatures had a mostly dampening influence on forecast dengue cases, consistent with some findings in the literature (Morin et al. 2013); and the breeding percentage, reflecting the preponderance of *Ae. aegypti*, was mostly positively correlated with forecast dengue incidence. Dengue incidence over the next 4–5 weeks was positively associated with high levels of absolute humidity over the last month and negatively associated with high humidity 15–20 weeks previously.

The forecasts at various time points in Singapore's record-breaking 2013 epidemic, in which 22,170 cases were reported, are presented in Figure 5. Early in the epidemic (Figure 5A), the model forecast was of a mild rise, which was exceeded by the actual epidemic. By February (Figure 5B), the forecast was an almost perfect match to the data. At the end of April (Figure 5C), the forecast was of a decline, but the range of possible scenarios (the 95% interval for the forecast) included the subsequently observed peak at ~800 cases/week. The end of the epidemic, starting in July, was also successfully forecasted. Overall, the model predicted a slightly more rapid end to, and smaller size of, the epidemic than that which occurred.

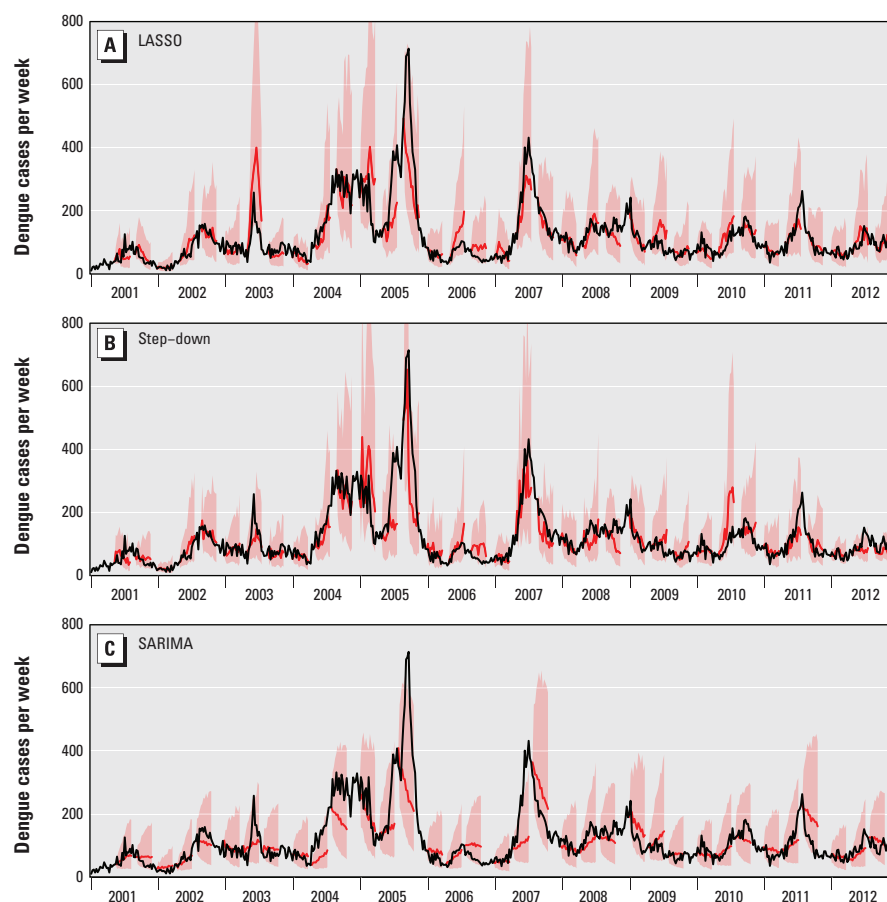
## Discussion

Penalized regression, of which LASSO is the most prominent methodology, is an idea that dates back to the 1990s (Tibshirani 2011) and over the last decade has led to substantial theoretical and methodological advances in “small  $n$  large  $p$ ” problems in which the number of observations is smaller than the number of potential predictors. By optimizing a combination of model goodness of fit (via the likelihood) and model complexity (via a penalty that grows with the size of the parameters in the model), and using cross-validation to identify the optimal penalty term, penalized regression can simultaneously handle both model building and parameter estimation, avoid over-fitting, and improve out-of-sample predictive accuracy (Hoerl and Kennard 1970; Zhao and Yu 2006). By shrinking regression coefficients of less importance to or towards zero and thereby retaining only the most important predictors, LASSO is able to obtain

good interpretability and stability (Zhao and Yu 2006).

Although LASSO is a well-established methodology that is frequently used in bioinformatics and other big-data applications (González-Recio et al. 2009; Shi et al. 2007; Wu et al. 2009), there are few applications of this method in epidemiology (for instance, Walter and Tiemeier 2009) or in neglected tropical diseases. In this paper, we report the use of LASSO in forecasting an endemic and high-burden disease—dengue—in Singapore, which, by virtue of its location near the equator and concomitant lack of seasonality, experiences unpredictable outbreaks above the usual endemic level. The LASSO methodology has several advantages over traditional approaches: *a*) Model selection is rapid (approximately 2 min), automatic and objective, in contrast to more laborious and sometimes subjective approaches such as the Hosmer and Lemeshow (2000) purposeful approach to model building, or common

approaches such as forward selection using  $p$ -values (Grechanovsky and Pinsker 1995). *b*) Tautologically, by selecting the model complexity using cross-validation to optimize predictive performance, predictive performance of the routine is optimized; thus, the task of making better and more accurate forecasts is reduced to that of finding potentially informative covariates. *c*) LASSO allows large numbers of possible predictors to be considered without prejudicing the accuracy of the routine because nonpredictive variables obtain zero coefficients for optimal values of the penalty term and thereby drop out of the final model. *d*) By using distinct models for 1-week forecasts, 2-week forecasts, and so on, the variables used and the values of their associated coefficients can be tailored to the specific requirements of forecasting different lengths of time into the future. Recent dengue cases (over the last 4 weeks) are important in forecasting for 1–12 weeks, and average temperature is not useful in short-term forecasts



**Figure 3.** Comparison of dengue forecast from 2001 to 2012. Dengue forecast using (A) LASSO, (B) step-down linear regression and (C) SARIMA. For the LASSO, step-down linear regression and SARIMA methods, we selected all of the data (2001 to 2012) except the year being forecast to develop the model (to approximate the out-of-sample predictive performance in the other results). In each plot, the red lines represent model-based point estimates, and the pink contours represent the corresponding 95% prediction intervals. Each segment of predicted data (i.e., each pink and red region) represents the estimates from one 12-week forecast made at a previous point in time. Forecasts at other times are suppressed to simplify the figure but are presented in Video Files S1, S2, and S3.

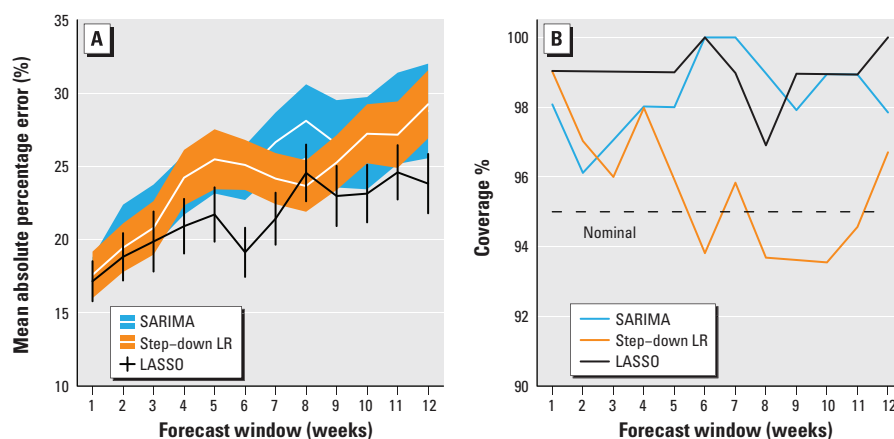
(1–3 weeks), although it is for longer forecasts (4–12 weeks); these examples highlight the need for separate submodels for different forecast windows. Having distinct submodels also obviates the need to forecast future values of the predictors, which would be the case if a single model for 1 week ahead were used and then iterated to obtain longer-term forecasts. This approach led to high accuracy for both immediate (next week, MAPE 17%) and long-term (3 months, MAPE 24%) predictions. Although the forecast accuracy degraded as the forecast window was extended (see Figure 4), this degradation was surprisingly slight, and we were able to predict the large outbreaks of both 2013 and 2014 over 10 weeks in advance, giving advance warning to allow operations to be rolled out. We restricted the forecast window to 12 weeks to avoid the increased level of inaccuracy that accompanies long-term projection and because short- (several weeks) and medium-term (several months) projections are the most useful for local planning purposes.

There are, however, some limitations to this approach. The largest of these limitations is that although very good predictive accuracy can be achieved, the 12 models built using the LASSO method are not amenable to interpretation because they were constructed for their predictive accuracy, not to explain the etiology of outbreaks. In particular, attempts to explain to stakeholders why the model forecast a large epidemic in 2013 were hindered by the numerous covariates acting at different lags. Interpretation is increasingly difficult at longer forecast windows. For example, ~60 predictors out of the complete set of > 200 were selected for the 12-week-ahead submodel. Among these 60 variables, the same covariate was often selected at different lags and frequently was selected with differently signed coefficients at those different lags. The complexity needed for good forecasts reflects the multitude of factors operating on the vector and virus–vector interactions. One plausible way to reduce the apparent complexity would be to combine our approach with a mechanistic model of drivers of the mosquito life cycle, for instance, via the Focks model (Focks et al. 1995), with output from the mechanistic model replacing some or all of the environmental drivers in the statistical model. The variables we used include meteorological data, case data, vector surveillance data, and human population data. Other relevant indicators of risk, particularly on circulating serotypes, genotypes, and evidence on herd immunity via occasional sero-epidemiological surveys, may subsequently be incorporated. However, because comprehensive analysis of genotype and serotype of dengue cases in Singapore began in approximately 2006, and because testing protocols have evolved since that time,

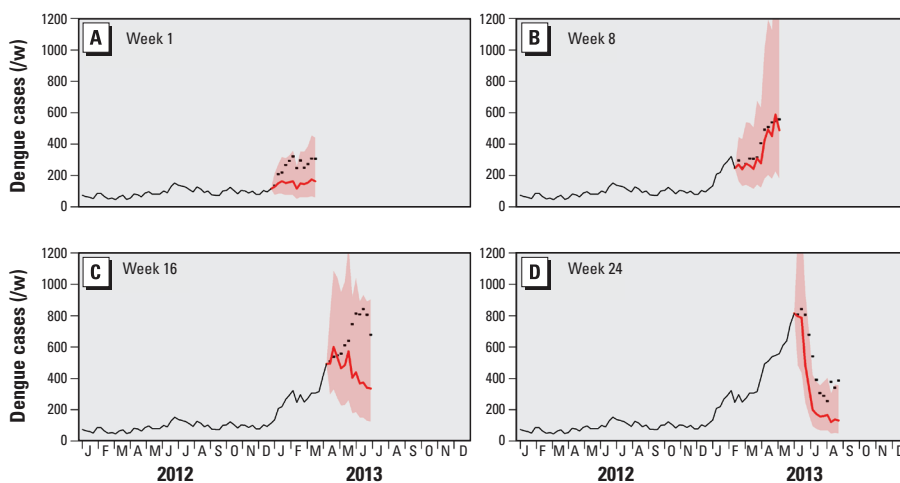
we need to explore the best way to incorporate these sources of information.

The forecasting tool described in this paper has become an integral part of Singapore's dengue control program. The final model is embedded in a “real-time” schedule, with data (at present) being updated weekly and predictions sent out to our operational partners (examples of the forecasts used in the 2013 outbreak are shown in Figure 5), including the Ministry of Health and the Environmental Public Health Operations Department of the NEA. During the 2013 epidemic, our forecasts helped guide hospital bed management and public health interventions, including

preemptive source reduction measures, recruitment of ground staff, and education campaigns. In late March 2013, our models forecasted an earlier-than-usual increase in dengue cases in June 2013, which could potentially peak at 800 cases/week. Specifically, the forecast predicted a peak in case count of 863 during the 26th week of 2013, which is very close to the observed number of cases, which peaked at 842 cases/week during the 25th week. In addition to aiding with resource planning, this forecast also facilitated early risk communication to the public and the advanced launch of Dengue Campaign in April, 2 months ahead of its traditional June launch.



**Figure 4.** Comparison of model performance among SARIMA, step-down linear regression (LR) and LASSO using mean absolute percentage error (MAPE) and coverage of nominal 95% intervals. (A) MAPE comparison among LASSO, step-down linear regression and SARIMA over 1- to 12-week forecast windows. MAPE with 95% confidence intervals for LASSO, step-down linear regression and SARIMA are represented by vertical bars, filled orange polygons, and filled blue polygons, respectively. The MAPE degraded more slowly over time under the LASSO approach, with an increase from 17% error forecasting at 1 week to 24% error forecasting at 3 months, although the LASSO approach yielded comparable accuracy to those of step-down linear regression and SARIMA for the first 2 weeks. (B) Coverage of LASSO, step-down LR and SARIMA. For each forecast window period, the percentage coverage was calculated using the number of observations that fell within the 95% interval derived by overlaying the estimated error distribution on top of the forecast. The dashed line represents the nominal coverage of 95%.



**Figure 5.** Illustration of model forecast in 2013 at weeks 1 (A), 8 (B), 16 (C) and 24 (D) using the LASSO method. In each panel, the x-axis represents time (2012/2013), and the y-axis represents the weekly number of dengue cases. The black lines indicate observed cases, the black dots indicate predicted (forecasted) cases, and the pink contours indicate the 95% intervals for the forecasts.



## Conclusion

Future work will automate the data-handling process so that predictions can be generated and posted online without the routine being rerun manually; such automation will also allow the forecast to be made daily and hence to be genuinely in “real time.” Extending the forecasts beyond 12 weeks may be challenging because some of the key drivers, such as local weather conditions, may have a short-term but strong effect on dengue that requires integrating the predictive model with weather forecasting models, where long-term forecasts may not be readily available.

## REFERENCES

- Althouse BM, Ng YY, Cummings DAT. 2011. Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis* 5:e1258, doi:10.1371/journal.pntd.0001258.
- Armstrong JS. 1985. Long-Range Forecasting: From Crystal Ball to Computer. 2nd ed. New York:Wiley-Interscience.
- Baguelin M, Hoek AJV, Jit M, Flasche S, White PJ, Edmunds WJ. 2010. Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. *Vaccine* 28:2370–2384, doi:10.1016/j.vaccine.2010.01.002.
- Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. 2013. The global distribution and burden of dengue. *Nature* 496:504–507, doi:10.1038/nature12060.
- Carrasco LR, Lee LK, Lee VJ, Ooi EE, Shepard DS, Thein TL, et al. 2011. Economic impact of dengue illness and the cost-effectiveness of future vaccination programs in Singapore. *PLoS Negl Trop Dis* 5:e1426, doi:10.1371/journal.pntd.0001426.
- Chatfield C. 2013. The Analysis of Time Series: An Introduction. 6th ed. Boca Raton, FL:Chapman & Hall/CRC Press.
- Coudeville L, Garnett GP. 2012. Transmission dynamics of the four dengue serotypes in southern Vietnam and the potential impact of vaccination. *PLoS One* 7:e51244, doi:10.1371/journal.pone.0051244.
- Cummings DAT, Iamsirithaworn S, Lessler JT, McDermott A, Prasanthong R, Nisalak A, et al. 2009. The impact of the demographic transition on dengue in Thailand: insights from a statistical analysis and mathematical modeling. *PLoS Med* 6:e1000139, doi:10.1371/journal.pmed.1000139.
- Department of Statistics, Singapore. 2015. Homepage. Available: <http://www.singstat.gov.sg/> [accessed 1 January 2015].
- Douglas DL, DeRoek DA, Mahoney RT, Wichmann O. 2013. Will dengue vaccines be used in the public sector and if so, how? Findings from an 8-country survey of policymakers and opinion leaders. *PLoS Negl Trop Dis* 7:e2127, doi:10.1371/journal.pntd.0002127.
- Earnest A, Tan SB, Wilder-Smith A, Machin D. 2012. Comparing statistical models to predict dengue fever notifications. *Comput Math Methods Med* 2012:e758674, doi:10.1155/2012/758674.
- Egger JR, Ooi EE, Kelly DW, Woolhouse ME, Davies CR, Coleman PG. 2008. Reconstructing historical changes in the force of infection of dengue fever in Singapore: implications for surveillance and control. *Bull World Health Organ* 86:187–196.
- Flores BE. 1986. A pragmatic view of accuracy measurement in forecasting. *Omega* 14:93–98, doi:10.1016/0305-0483(86)90013-7.
- Focks DA, Daniels E, Haile DG, Keesling JE. 1995. A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. *Am J Trop Med Hyg* 53:489–506.
- Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.
- González-Recio O, de Maturana EL, Vega AT, Engelman CD, Broman KW. 2009. Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model. *BMC Proc* 3(suppl 7):S63, doi:10.1186/1753-6561-3-S7-S63.
- Grechanovsky E, Pinsky I. 1995. Conditional p-values for the F-statistic in a forward selection procedure. *Comput Stat Data Anal* 20:239–263, doi:10.1016/0167-9473(94)00047-M.
- Halstead SB. 1999. Is there an inapparent dengue explosion? *Lancet* 353:1100–1101, doi:10.1016/S0140-6736(05)76460-6.
- Halstead SB. 2012. Dengue vaccine development: a 75% solution? *Lancet* 380:1535–1536, doi:10.1016/S0140-6736(12)61510-4.
- Hii YL, Rocklöv J, Wall S, Ng LC, Tang CS, Ng N. 2012a. Optimal lead time for dengue forecast. *PLoS Negl Trop Dis* 6:e1848, doi:10.1371/journal.pntd.0001848.
- Hii YL, Zhu H, Ng N, Ng LC, Rocklöv J. 2012b. Forecast of dengue incidence using temperature and rainfall. *PLoS Negl Trop Dis* 6:e1908, doi:10.1371/journal.pntd.0001908.
- Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12:55–67, doi:10.1080/00401706.1970.10488634.
- Hosmer DW, Lemeshow S. 2000. Applied Logistic Regression. 2nd ed. New York:Wiley-Interscience Publication.
- Hughes H, Britton NF. 2013. Modelling the use of Wolbachia to control dengue fever transmission. *Bull Math Biol* 75:796–818, doi:10.1007/s11538-013-9835-4.
- Lee C, Vythilingam I, Chong CS, Abdul Razak MA, Tan CH, Liew C, et al. 2013. Gravitrap for management of dengue clusters in Singapore. *Am J Trop Med Hyg* 88:888–892, doi:10.4269/ajtmh.12-0329.
- Lee KS, Lai YL, Lo S, Barkham T, Aw P, Ooi PL, et al. 2010. Dengue virus surveillance for early warning, Singapore. *Emerg Infect Dis* 16:847–849, doi:10.3201/eid1605.091006.
- Lee KS, Lo S, Tan SSY, Chua R, Tan LK, Xu H, et al. 2012. Dengue virus surveillance in Singapore reveals high viral diversity through multiple introductions and *in situ* evolution. *Infect Genet Evol* 12:77–85, doi:10.1016/j.meegid.2011.10.012.
- Lim SP, Wang QY, Noble CG, Chen YL, Dong H, Zou B, et al. 2013. Ten years of dengue drug discovery: progress and prospects. *Antiviral Res* 100:500–519, doi:10.1016/j.antiviral.2013.09.013.
- Mahalingam S, Herring BL, Halstead SB. 2013. Call to action for dengue vaccine failure. *Emerg Infect Dis* 19:1335–1337, doi:10.3201/eid1908.121864.
- MOH (Ministry of Health, Singapore). 2012. Communicable Diseases Surveillance in Singapore 2011. Available: [http://www.moh.gov.sg/content/moh\\_web/home/Publications/Reports/2012/\\_communicable\\_diseasesurveillanceinsingapore2011.html](http://www.moh.gov.sg/content/moh_web/home/Publications/Reports/2012/_communicable_diseasesurveillanceinsingapore2011.html) [accessed 23 June 2014].
- Morin CW, Comrie AC, Ernst K. 2013. Climate and dengue transmission: evidence and implications. *Environ Health Perspect* 121:1264–1272, doi:10.1289/ehp.1306556.
- Murphy BR, Whitehead SS. 2011. Immune response to dengue virus and prospects for a vaccine. *Annu Rev Immunol* 29:587–619, doi:10.1146/annurev-immunol-031210-101315.
- Ng LC, Tan LK, Tan CH, Tan SSY, Hapuarachchi HC, Pok KY, et al. 2009. Entomologic and virologic investigation of Chikungunya, Singapore. *Emerg Infect Dis* 15:1243–1249, doi:10.3201/eid1508.081486.
- Ong JBS, Chen MIC, Cook AR, Lee HC, Lee VJ, Lin RTP, et al. 2010. Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PLoS One* 5:e10036, doi:10.1371/journal.pone.0010036.
- Peel MC, Finlayson BL, McMahon TA. 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci* 11:1633–1644, doi:10.5194/hess-11-1633-2007.
- R Core Team. 2014. R: A Language and Environment for Statistical Computing. Vienna, Austria:R Foundation for Statistical Computing. Available: <http://www.R-project.org> [accessed 1 January 2014].
- Rathore APS, Paradkar PN, Watanabe S, Tan KH, Sung C, Connolly JE, et al. 2011. Celgosivir treatment misfolds dengue virus NS1 protein, induces cellular pro-survival genes and protects against lethal challenge mouse model. *Antiviral Res* 92:453–460, doi:10.1016/j.antiviral.2011.10.002.
- Rigau-Pérez JG, Clark GG, Gubler DJ, Reiter P, Sanders EJ, Vorndam AV. 1998. Dengue and dengue haemorrhagic fever. *Lancet* 352:971–977, doi:10.1016/S0140-6736(97)12483-7.
- Rosen L, Shroyer DA, Tesh RB, Freier JE, Lien JC. 1983. Transovarial transmission of dengue viruses by mosquitoes: *Aedes albopictus* and *Aedes aegypti*. *Am J Trop Med Hyg* 32:1108–1119.
- Sabchareon A, Wallace D, Sirivichayakul C, Limkittikul K, Chanthavanich P, Suvannadabba S, et al. 2012. Protective efficacy of the recombinant, live-attenuated, CYD tetravalent dengue vaccine in Thai schoolchildren: a randomised, controlled phase 2b trial. *Lancet* 380:1559–1567, doi:10.1016/S0140-6736(12)61428-7.
- Shi W, Lee KE, Wahba G. 2007. Detecting disease-causing genes by LASSO-Patternsearch algorithm. *BMC Proc* 1:S60–S64.
- Thu HM, Aye KM, Thein S. 1998. The effect of temperature and humidity on dengue virus propagation in *Aedes aegypti* mosquitoes. *Southeast Asian J Trop Med Public Health* 29:280–284.
- Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol* 58:267–288.
- Tibshirani R. 2011. Regression shrinkage and selection via the lasso: a retrospective. *J R Stat Soc Series B Stat Methodol* 73:273–282, doi:10.1111/j.1467-9868.2011.00771.x.
- Verbesselt J, Hyndman R, Newnham G, Culvenor D. 2010. Detecting trend and seasonal changes in satellite image time series. *Remote Sens Environ* 114:106–115, doi:10.1016/j.rse.2009.08.014.
- Walter S, Tiemeier H. 2009. Variable selection: current practice in epidemiological studies. *Eur J Epidemiol* 24:733–736, doi:10.1007/s10654-009-9411-2.
- Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25:714–721, doi:10.1093/bioinformatics/btp041.
- Xu HY, Fu X, Lee LKH, Ma S, Goh KT, Wong J, et al. 2014. Statistical modeling reveals the effect of absolute humidity on dengue in Singapore. *PLoS Negl Trop Dis* 8:e2805, doi:10.1371/journal.pntd.0002805.
- Zhao P, Yu B. 2006. On model selection consistency of lasso. *J Mach Learn Res* 7:2541–2563.