Journal of Biomedical Informatics 81 (2018) 16–30

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison

Yirong Chen[a], Collins Wenhan Chu[b], Mark I.C. Chen[a,c], Alex R. Cook[a,*]

[a] Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Tahir Foundation Building, 12 Science Drive 2, 117549, Singapore

[b] Genome Institute of Singapore, 60 Biopolis Street, Genome, 138672, Singapore

[c] Department of Clinical Epidemiology, Communicable Disease Centre, Tan Tock Seng Hospital, Singapore, Moulmein Road, 308433, Singapore

ARTICLE INFO

Keywords:
LASSO
Endemic infectious disease
Real time forecast

ABSTRACT

Introduction: Accurate and timely prediction for endemic infectious diseases is vital for public health agencies to plan and carry out any control methods at an early stage of disease outbreaks. Climatic variables has been identified as important predictors in models for infectious disease forecasts. Various approaches have been proposed in the literature to produce accurate and timely predictions and potentially improve public health response.

Methods: We assessed how the machine learning LASSO method may be useful in providing useful forecasts for different pathogens in countries with different climates. Separate LASSO models were constructed for different disease/country/forecast window with different model complexity by including different sets of predictors to assess the importance of different predictors under various conditions.

Results: There was a more apparent cyclicity for both climatic variables and incidence in regions further away from the equator. For most diseases, predictions made beyond 4 weeks ahead were increasingly discrepant from the actual scenario. Prediction models were more accurate in capturing the outbreak but less sensitive to predict the outbreak size. In different situations, climatic variables have different levels of importance in prediction accuracy.

Conclusions: For LASSO models used for prediction, including different sets of predictors has varying effect in different situations. Short term predictions generally perform better than longer term predictions, suggesting public health agencies may need the capacity to respond at short-notice to early warnings.

1. Introduction

Outbreaks such as those caused by the Severe Acute Respiratory Syndrome Coronavirus (SARS CoV), the influenza A(H1N1)pdm09 pandemic of 2009, and more recently the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), Ebola virus and Zika virus have demonstrated the high potential risk of emerging and re-emerging infectious diseases to spread within and between countries [1–5]. These in turn cause increasing challenges for public health systems, including the increasing burden of infectious disease, and the need to build a surveillance and response system that is able to identify newly emerging disease rapidly, both regionally and internationally which calls for international collaboration, and the need for drug and vaccine research and production [6–8]. While the response to endemic diseases may be less urgent, the burden caused by pathogens such as influenza or malaria is high [9–11], and due to their endemicity, many countries have long standing surveillance systems to track outbreaks and guide response, from vector control to hospital bed utilization [12–15]. Early warning systems aiming to predict epidemics as soon as possible can allow control methods to be carried out rapidly and increase their chance of success [16,17]. To do so, decision makers need to be able to make accurate forecasts of incidence and to automate this forecasting process based on routinely collected notification data [18]. If accurate forecasts were available in both the near and far future, effective policies could then be targeted to the expected future needs. Existing approaches to real-time forecasting include generalized linear regression, seasonal autoregressive integrated moving average (SARIMA) model or a simpler ARIMA form of it, phenomenological models like the logistic growth model and Richards model, and mechanistic models like the SIR models [19–24]. Often such approaches involve the challenge of integrating environmental factors including temperature, humidity and rainfall, which may influence pathogen transmission directly or affect

* Corresponding author.
E-mail address: alex.richard.cook@gmail.com (A.R. Cook).

https://doi.org/10.1016/j.jbi.2018.02.014
Received 11 September 2017; Received in revised form 19 January 2018; Accepted 24 February 2018
Available online 27 February 2018
1532-0464/

the vector activities (for vector borne diseases), especially in temperate regions [25–29]. For instance, influenza virus is more transmissible in low temperature and low humidity conditions [30,31], while the primary vector of dengue, the yellow fever mosquito *Aedes aegypti*, favors higher temperature [32,33]. The availability of real time data-streams on seasonal variation and climatic variability therefore holds the potential to lead to more accurate prediction algorithms, potentially improving public health response.

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a machine learning method that can find patterns within large datasets while avoiding the problem of over-fitting [34]. Estimation and variable selection are simultaneously carried out using the LASSO method, and as such it is commonly used in studies in fields with large numbers of explanatory variables to reduce the variable space. This algorithm trades off model accuracy with model parsimony by introducing a penalty term into the objective function (which in standard linear regression is the sum of squares of residuals). The penalty term can, for linear regression models, be made equivalent to a constraint on the sum of the absolute parameter coefficients. This constraint imposed by LASSO regression has the effect of shrinking some estimated coefficients towards zero, which may help reduce biases caused by separation in some forms of regression [35], while simultaneously producing some parameter estimates that are exactly 0, so that the covariate associated with this coefficient is not associated with the outcome variable in that model. The optimal balance between model accuracy and complexity is typically obtained through cross-validation: repeatedly partitioning the data into training and validation sets, varying the degree of penalty, optimizing the regression parameters for each penalty value, then selecting the penalty that minimizes out of sample predictive accuracy. Computationally efficient methods to explore the penalty and parameter space exist [36], making it feasible to use LASSO as part of a 'real-time' forecasting pipeline for routinely collected health data such as infectious disease notifications. This computational speed allows the forecasting to adapt to changes in the underlying disease dynamics by permitting refitting of the model each time new data are reported, which may be important for diseases in which the severity changes between outbreaks, such as influenza [37]. Forecasts at different time horizons can be obtained through splicing together separate LASSO models, each trained on the data available at the time of the forecast, but tailored to predict at different windows into the future.

The LASSO method has previously been used in dengue outbreak prediction in Singapore, where it is now routinely used to guide vector control policy [38]. The objective of this paper is to apply the LASSO method to infectious disease forecasting and assess more generally in which situations LASSO models will provide useful forecasts. Unlike conventional use of the LASSO method to variable selection, the primary interest of our application of the LASSO-based method on infectious disease data is to make forecast of incidences in the future, rather than to identify the potential risk factors to explain the epidemics of these infectious diseases.

In particular, we assess for diseases with different transmission modalities, in different climatic zones, how accurate short to medium time forecasts can be, and what data streams are necessary for accurate forecasts. We apply the method to four countries from different latitudes—Japan, Taiwan, Thailand and Singapore—to cover temperate, sub-tropical and tropical settings.

2. Methods

2.1. Sources of data

Four representative countries with distinct climates were selected for analysis based on Köppen-Geiger climate classification [39] – Japan with humid continental and subtropical climate; Taiwan with humid subtropical and oceanic climate; Thailand with tropical wet and savannah climate and Singapore with tropical rainforest climate. Four representative infectious diseases were included in the study: two mosquito-borne infections (Dengue and Malaria) and two infections that spread from person to person (Hand Foot and Mouth Disease (HFMD) and Chickenpox). For all four pathogens, a relationship has previously been found between incidence and climatic variables [40–42] or for there to be a seasonality to incidence [43]. Not all four pathogens were considered for each country: some are not present in each country while others are not captured in routine infectious disease surveillance systems.

The notified numbers of chickenpox, HFMD cases in Japan were collected by the National Institute of Infectious Diseases (NIID) [44]. Both were reported as average cases per week per sentinel reporting, to accommodate varying reporting rates. We extracted weekly data from 2001 to 2012.

Monthly reported cases of chickenpox, dengue, and malaria in Thailand for the period 2003–2013 were obtained from the Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Thailand [45]. The number of incident cases were collected from government hospitals, public health offices and health centers by the National Disease Surveillance [46] and were reported online.

Ministry of Health, Singapore, actively monitors and publishes the incidence of dengue and HFMD in Singapore, both being notifiable diseases. Weekly number of incidences for the period 2003 and 2014 were obtained from the Weekly Infectious Diseases Bulletin [47].

Weekly number of dengue cases from 2003 to 2014 were extracted from Taiwan National Infectious Disease Statistics System [48]. Both indigenous and imported cases were included in the count.

Epidemiological week as per US Centers for Disease Control and Prevention was used in our analysis using the EpiWeek package in R [49].

Climatic data for Taiwan, Thailand and Singapore were obtained from the Weather Underground [50] which documented among other variables, historical temperature, humidity, sea level pressure, and visibility. Only temperature (daily highest, average and lowest) and relative humidity (daily highest, average and lowest) were used in our models due to insufficient historical data of other climatic variables. Climatic data for Japan were obtained from the Japan Meteorological Agency [51], which provides and archives various weather information. Weekly mean temperature, relative humidity and rainfall information were used in our model. For all locations, the weather data at the capital (Tokyo, Taipei, Bangkok and Singapore) was used to represent overall national weather.

2.2. Statistical analysis

Wavelet analyses were done to explore periodicity of all endemic diseases and climatic variables in four countries. The wavelet approach was based on a wavelet function which analyses locality in time and frequency [52]. Wavelet transformation ($W_t(s)$) as the convolution of the time series x_t with Morlet function $\psi_0(\eta)$ at scale s was conducted:

$$W_t(s) = \sum_{k=0}^{T-1} \hat{x}_k \hat{\psi}^*(s\omega_k) e^{i\omega_k n\delta_t}$$

where $\hat{x}_k$ is the discrete Fourier transform of x_t: $\hat{x}_k = \frac{1}{T}\sum_{t=0}^{T-1} x_t e^{-2\pi ikt/T}$, $k = 0,...,T-1$ the frequency index, and $\hat{\psi}(s\omega)$ the Fourier transform of Morlet function: $\hat{\psi}(s\omega) = \pi^{-\frac{1}{4}} H(\omega) e^{-(s\omega-\omega_0)^2/2}$, where ω_0 refers to the nondimensional frequency and is set to 6 to satisfy the admissibility condition [53].

The wavelet transformation $W_t(s)$ can be divided into amplitude, $|W_t(s)|$, and phase, $\tan^{-1}[\Im\{W_n(s)\}/\Re\{W_n(s)\}]$, where $\Re\{W_n(s)\}$ is the real part of the transform and $\Im\{W_n(s)\}$ the imaginary part. The wavelet power spectrum is defined as $|W_t(s)|^2$ [54].