Can we identify ‘bellwether’ states with respect to syphilis incidence?

Mauricio Monsalve*, Sean Tolentino, Sriram Pemmaraju and Philip Polgreen
Computational Epidemiology Group, The University of Iowa, Iowa City, IA, USA

Objective
(i) To forecast syphilis cases per state in the United States to support early containment of outbreaks. (ii) For each state, to determine which states are most correlated, to find ‘bellwether’ states to inform surveillance efforts. (iii) To determine a small collection of states whose syphilis incidence patterns are most closely correlated with all the states.

Introduction
The time series of syphilis cases has been studied at the country and state level at the yearly basis (1, 2), and it has been found that syphilis has a periodicity of approximately 10 years (2). However, to inform prevention efforts, it is important to understand the short-term dynamics of disease activity.

Methods
We used data from the MMWR. It contains weekly syphilis counts per state. We consider the time period from 1995 to 2009. We removed week 53 when present, due to inconsistencies in reporting. We considered 53 locations: the 50 states plus Puerto Rico, and the cities of New York City and Washington DC. To predict disease activity in each state, we constructed a series of linear lagged regression models that used several states as covariates. To benchmark our models, we constructed a basic ARIMA model with one autocorrelation term. All the models were constructed to forecast 4 weeks in advance. Prediction at week t was performed by fitting the models using all past data prior to week t.

To identify bellwether states, we proceeded as follows. First, we repeatedly fitted 2-covariate models to forecast each state and obtained the top 5 most frequent bellwether states for each state. Then, we obtained the most frequent bellwether states from the above lists.

Results
We found that forecasting states using less than 10 states as covariates is better than using more or the state itself as covariate (ARIMA), as shown in Fig. 1. An example of out of sample prediction is shown in Fig. 2, for New York City.

Conclusions
We also found that the 10 most frequent states in models with two covariates are California, Virginia, Florida, New York City (treated as state), Alabama, Ohio, Tennessee, North Carolina, New Hampshire and New Mexico. The first 5 are covariates of 40 states, and the amount increases to 50 when adding the later.

Keywords
Syphilis time series; forecasting; disease surveillance

References

*Mauricio Monsalve
E-mail: mauricio-monsalve@uiowa.edu