Syndromic Prediction Power: Comparing Covariates and Baselines

John Copeland[1], Gabriel Rainisch[1], Jerry Tokars[1], Howard Burkom[2], Nancy Grady[3], Roseanne English[1]

[1] CDC National Center for Public Health Informatics, Division of Emergency Preparedness and Response
[2]Johns Hopkins University, Applied Physics Lab, National Security Technology Department
[3] Science Applications International Corporation, Department of Adaptive Enterprise Solutions

OBJECTIVE

This paper compares the prediction accuracy of regression models with different covariates and baseline periods, using a subset of data from CDC's BioSense initiative. Accurate predictions are needed to achieve sensitivity at practical false alarm rates in anomaly detection for biosurveillance.

BACKGROUND

The eleven syndrome classifications for clinical data records monitored by BioSense include rare events such as death or lymphadenitis and also common occurrences such as respiratory infections. BioSense currently uses two statistical methods for prediction and alerting with respect to the eleven syndromes. These are a modified CUSUM; and small area regression and testing (SMART), described by Ken Kleinman [1]. At the inception of BioSense, these prediction methods were implemented as one-model-fits-all, and they remain largely unmodified. An evaluation of the predictive value of these methods is required. The SMART method, as used in BioSense, uses long-term data. As covariate predictors, day-of-week, a holiday indicator, day after holiday, and sine/cosine seasonality variables are used. Lengthy, stable historical data is not always available in BioSense data sources, and this obstacle is expected to grow as data sources are added. We wish to test regression methods of surveillance that use shorter time periods, and different sets of predictors.

METHODS

The holiday indicator variable was reconsidered. Rather than using calendar holidays, the models presented in this paper define a holiday as a day off for federal employees. Using this definition of a holiday, four predictor variables were considered in Poisson regression models: day-of-week, holiday, total daily health facility visits, and a trend variable (date). Models using different combinations of these covariates were compared. The test data for this paper consisted of daily military clinical visit count data from the Department of Defense (DoD) for the entire state of Texas, from 9/1/2004 to 3/31/2006. Analyses were performed at the treatment facility level using SAS PROC GENMOD. Predictions were computed for syndromic counts of lymphadenitis visits, to represent rare syndromes, respiratory visits, to represent common syndromes, and gastrointestinal visits, to represent syndromes in between. For each syndrome and facility, baseline windows of 28, 56, and 112 days were used to make one-day-ahead visit count predictions. For each forecast, the baseline period was shifted forward by one day for model inferences based on the immediately preceding time window. Thus, 450-550 days of predicted counts for each facility and syndrome were computed and the absolute value of the residual was used to measure the prediction accuracy of the model.

RESULTS

With 97 available facilities, non-convergence allowed the use of only 18 facilities for respiratory models, 17 facilities for gastrointestinal models, and 10 facilities for lymphadenitis models. Mean syndrome counts, respectively, were 71 on weekdays and 20 on weekends/holidays, 20 on weekdays and 8 on weekends and holidays, and 1.09 on weekdays and 0.25 on weekends and holidays. Models containing total facility visits resulted in more accurate syndrome count predictions than models without total visits. This improvement was seen regardless of what other covariates were used in the model, and regardless of syndrome or baseline period. For the (highly seasonal) respiratory syndrome, the shorter 28-day baseline resulted in more accurate predictions than the longer 56 or 112-day baselines. For the rarer lymphadenitis, which showed no seasonality, the opposite is true: A 112-day baseline yielded the most accurate predictions.

Table. Median Absolute Residuals by Syndrome and Regression Model (28-Day Baseline).

	Lymp	Gast	Resp
1. DOW, holiday	0.4646	2.500	9.012
2. DOW, holiday, trend	0.4715	2.458	8.756
3. DOW, holiday, visits	0.4497	2.234	7.619
4. DOW, holiday, trend, visits	0.4490	2.201	7.421
5. DOW, trend, visits	0.4459	2.167	7.390

CONCLUSIONS

We compared the prediction accuracy of four predictor variables and three baseline periods in Poisson regression models. Among the strongest predictors for counts of all tested syndromes was the total number of facility visits for a particular day. These results suggest that accounting for total visits will produce a substantial improvement in prediction accuracy. The rationale for using this covariate is that the total facility count is expected to display many of the same system-related effects (excluded from other covariates) as the syndromic count, but should show relatively little effect from a syndrome-specific outbreak. For well represented and highly seasonal syndromes such as respiratory, the results suggest that a shorter baseline period produces a more accurate prediction. This finding is not unexpected if the baseline is long enough to modulate daily fluctuations (and 28 days seems minimal for this purpose), as the shorter period may better reflect recent trends than will a longer period. The results suggest that the opposite is true for sparse non-seasonal syndromes, as trends are less prevalent and more data is needed for sensible inference.

More research needs to be done to determine the common causes of non-convergence in syndromic count data, and to provide evidence for data aggregation decisions to avoid this problem. Furthermore, a primary aim of syndromic surveillance is anomaly detection, and the predictions evaluated here are needed for detection decisions, but statistical alerting mechanisms are required for sensitivity at practical false alarm rates. Additional topics for further research include investigating additional syndromes and determining the effect on sensitivity and specificity for detecting clusters.

REFERENCES

[1] Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. Am J Epidemiol 2004;159:217-24

Further Information: John Copeland, JCopeland@cdc.gov