Comparison of Regression Models with Modified Time Series Methods for BioSurveillance

Jian Xing[1], Howard Burkom[2], Jerome Tokars[1]

[1]Centers for Disease Control and Prevention, [2]The Johns Hopkins Applied Physics Lab,

OBJECTIVE

To compare regression models with the modified C2 algorithm for analysis of time series data and real time outbreak detection.

BACKGROUND

Previous studies using BioSense System data showed that adjusting for total visits in regression models improves accuracy of predicted value calculations [1]; and modification of the EARS C2 algorithm, including adjusting for total facility visits and a longer baseline for stable thresholding, improves the accuracy of expected value calculation and overall performance [2]. A previous study of data aggregated from 1 metropolitan area showed that Poisson regression modeling performed better than C2; however, this comparison did not include the adjustments above [3]. Therefore, we compare regression model estimation methods with the modified EARS C2 method.

METHODS

The study data source was BioSense national hospital emergency department chief complaint data that included records from >400 facilities. These records were classified into the standard 11 BioSense syndrome groups. We present here preliminary analyses using respiratory syndrome counts during May 1, 2007 to April 30, 2008. We calculated expected values using the modified EARS C2 method with a sliding 28-day baseline and adjustment for total visits. We also used a linear regression model controlling for total visits, day of week, and 14-day time period (a separate indicator variable for each 14-day period); the model was run separately for each facility, with the expected value for each day in the study period calculated from regression coefficients using the previous 90 days of data, with a 2-day buffer. We compared mean absolute residuals during the full year; and stratifying by mean observed count during 3-months periods in the flu season (Jan-Mar 2008) and non-flu season (May-July 2007).

RESULTS

The mean observed count per facility per day was 19.95 (range 0.14 to 98.27). Over the full year, the mean absolute residual was lower for modified C2 than regression (3.52 vs. 3.76). During the non-flu season, residuals were lower for modified C2 overall and across all observed count categories (Table). During flu season, residuals were modestly lower for regression overall and at mean observed counts >20 per facility per day.

Table: Mean Absolute Residuals by Observed Count, Method, and Time Period.

	May-July, 2007			Jan-Mar, 2008		
Mean Observed Counts	% of total case	Regres-sion	Modified C2	% of total case	Regres-sion	Modified C2
Over all		3.32	3.02*		4.24**	4.33
0 to 5	17.09	1.16	1.00*	9.38	1.21	1.07*
6 to 10	20.49	2.47	2.22*	14.04	2.54	2.37*
11 to 20	33.25	3.44	3.07*	26.13	3.40	3.31*
21 to 30	15.43	4.41	3.99*	20.60	4.40**	4.49
31 to 40	9.54	5.42	5.18*	12.03	5.62**	5.89
41 to 50	3.09	5.98	5.64*	7.64	6.26**	6.61
51 to 60	0.28	5.75	6.07*	4.17	6.78**	7.18
61 to 70	0.28	8.31	7.39*	1.62	7.59**	7.76
> 70	0.56	9.55	9.40*	4.40	9.44**	10.72

* Modified C2 better
** Regression better

CONCLUSIONS

These preliminary results demonstrate that modified C2 produces lower residuals during non-flu season and for facilities with mean observed counts <20 per day during flu season. Further studies will examine additional syndromes, aggregation at the city level, Poisson regression modeling, and sensitivity to injected signals representing data effects of outbreaks.

REFERENCES

[1] Copeland, J, et al. Syndromic prediction power: comparing covariates and baselines. Adv Dis Surv 2007;2:46
[2] Xing J, Burkom H, Copeland J, Bloom S, Hutwagner L, Tokars J. Performance Characteristics of Control Chart Detection Methods. Adv Dis Surv 2007;4:122.
[3] Jackson M, et al. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Medical Informatics and Decision Making. 2007, 7:6.