The Predictive Accuracy of Non-Regression Data Analysis Methods

Jerome I. Tokars M.D.[1], Roseanne English M.S.[1], Steve Bloom M.S.[2]
BioSense Project, [1]Centers for Disease Control and Prevention, [2]Science Applications International Corporation

Objective: To compare the predictive accuracy of three non-regression methods in analysis of time series count data.

Background: Analysis of time series data requires accurate calculation of a predicted value. Non-regression methods such as the Early Aberration Reporting System (EARS) CuSum are computationally simple, but most do not adjust for day of week (DOW) or holiday. Alternately, regression methods require larger counts, more computer resources, and possibly longer baseline periods of data. As increasing volumes of data are reported and analyzed, the predictive accuracy of simpler methods should be assessed and optimized.

Methods: Data from Veterans Affairs (VA) and Department of Defense (DoD) outpatient facilities, as reported to the CDC BioSense system, from August 1, 2004 to December 30, 2005 from Virginia and Texas were analyzed. Data were aggregated to the 11 BioSense syndromes at the facility level. Facilities that rarely reported syndromic data were excluded. We used three methods to calculate predicted values: (1) EARS CuSum C1 (uses the mean of day-7 to day-1) and C2 (uses the mean of day-10 to day-3); (2) modified CuSum, where weekend days are compared with weekend days and weekdays with weekdays (calculations corresponding to C1 and C2 are designated W1 and W2); and (3) a DOW adjusted method (DOWA). W1 and W2 are used in BioSense for analysis of VA and DoD data aggregated on the state level. Bias in predicted value calculation was assessed by summing the residuals (observed-predicted) for each DOW. Predictive accuracy was measured by the absolute value of the residual. DOWA values were calculated as follows. Preliminary analyses showed that counts on Federal holidays were similar to those on Sundays; therefore, holidays were analyzed as if they were Sundays. Data from the 56 days before an index day were used. The mean count during the prior week ($mean_{week}$) and the mean for each DOW (7 values for $mean_{dow}$) were calculated. Using the m_{dow} values, the expected mean count during the prior week ($mean_{expect}$) was calculated. The DOWA predicted value was then = $m_{week}*m_{dow}/m_{expect}$.

Results: A total of 559,251 facility-syndrome-days were analyzed at 75 (56 VA and 19 DoD) Texas and 26 (19 VA and 17 DoD) Virginia facilities. The mean observed count per facility per day varied among the syndromes from <.01 for severe injury/death to 21.0 for respiratory. For all facilities and syndromes, the mean observed daily count was 3.95; counts were <=2 on 75% of days. There were minimal differences between the mean absolute residual for C1 (2.31) vs C2 (2.33); or for W1 (1.22) vs W2 (1.24); the mean absolute residual was 1.17 for DOWA. These differences were greater as the mean count increased: for a mean count of 45-55, the mean absolute residual was 20.1 for C1, 10.3 for W1, and 9.5 for DOWA. Bias among DOW was high using C1 (among the DOWs, mean residuals ranged from -3.0 to 1.9), lower for W1 (range -0.7 to 0.6), and lowest for DOWA (range -0.02 to 0.02).

Conclusions: For most facility-syndrome-day combinations, observed counts are low, making regression methods impractical. Compared with C1 or C2, the calculation of predicted value is much more accurate using W1 or W2, with a small additional improvement for DOWA, especially at high mean counts. Because of minimal bias among DOW, the DOWA method should have a similar chance of finding data anomalies on each DOW. These methods are computationally simple, always produce a predicted value, work with <=56 days of prior data, and unlike regression methods do not have convergence problems. Further evaluation should include assessment of moving averages >1 week, exclusion of outlier values from moving average calculations, comparisons of prediction accuracy with regression models, evaluations of sensitivity and specificity for cluster detection, adjustment for total number of visits, and assessment using other data sources and events. Non-regression methods may be a valuable component of a multi-method data analysis strategy.