Geographic Categorization Methods used in BioSense

Roseanne English[1], BS, Paul C. McMurray[2], MDS, Leslie Z. Sokolow[3], MSc, CGIS, Henry Rolka[1], MS, MPS, David Walker[1], MPH, John F. Quinn III[4], BS, Kenneth L. Cox[5], Col, USAF, MC, SFS

[1]BioIntelligence Center, Centers for Disease Control and Prevention; [2]Science Application International Corporation; [3]Northrop Grumman; [4]U.S. Department of Veterans Affairs; [5]Department of Defense

OBJECTIVE

Precise geographic location of health events is a challenging but critical component to determine the likely site of exposure for disease surveillance. This paper describes a method used by BioSense to develop and implement a reasonable set of rules in defining geographic locations of health events.

BACKGROUND

BioSense is a CDC initiative to promote situational awareness through summarizing, analyzing, and presenting health related event information. Among the data sources collected and analyzed through the BioSense application are the Department of Defense (DoD) and Department of Veterans Affairs (VA) ambulatory clinic care data. Clinical diagnoses and procedures are quantified, and analytic results are presented and categorized into 94 state and metropolitan areas.

BioSense originally categorized the geographic location of DoD data based on the zip code of the health facility visited and the VA data based on the patient residence zip code. While these geographic categorization rules work well within BioSense, it is prudent to identify and implement more robust rules so as to accurately categorize the geographic location of an event of potential interest.

BioSense has approached the use of an alternate zip code for geographic location purposes based on distance between the patient residence zip code and the health facility zip code. The season during which the health event occurred is also a factor taken into account.

METHODS

The distance in miles between the patient residence zip code and the health facility zip code was calculated and examined for each geographic area and health facility. Various reports and plots were generated to visualize the results. The plots revealed the distance distributions were highly skewed to right. An alternate standard deviation was computed using the Interquartile Range (IQR) as σ= (IQR/1.34898) [1]. A maximum “distance traveled” cutoff was computed for each health facility based on the alternate standard deviation. Visits with a mileage less than or equal to the maximum distance cutoff were categorized as local visits, whereas visits with a mileage greater than the maximum distance cutoff were categorized as remote visits. An alternate zip code was then used in the geographic categorization of records based on their local vs. remote categorization. Local visits were categorized into geographic areas based on the patient residence zip code, whereas remote visits were categorized into geographic areas based on the health facility zip code.

The Wilcoxon Rank Sum Test for comparing two independent samples was used to determine if the distance traveled to a given facility differed by season. Analyses were run to calculate the percent change in geographic location, per data source and season, when the location was defined based on the derived alternate zip code.

RESULTS

The Wilcoxon Rank Sum Test revealed that 45% of the DoD health facilities and 32% of the VA health facilities had seasonal differences. When applying the distance analysis rules, the geographic area was different in 8% of DoD visits and 17% of VA visits during the Winter season, and different in 9% of the DoD visits and 13% of the VA visits during the Summer season.

CONCLUSION

A reasonable set of rules based on distances can be used in defining a geographic location in BioSense data. The use of these rules augments BioSense’s capacity for presenting the likely site of exposure of health events.

REFERENCES

[1] Tukey, J.W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley).

Roseanne English, rxe1@cdc.gov