Emerging Health Threats Journal 2011, **4**:s49. **doi:** 10.3134/ehtj.10.049

www.eht-journal.org

ABSTRACT

A data simulation model using NRDM pharmaceutical sales counts

Jialan Que, and Fu-Chiang Tsui

RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
E-mail: jiq4@pitt.edu

Objective
This study proposes a simulation model to generate the daily counts of over-the-counter medication sales, such as thermometer sales from all ZIP code areas in a study region that include the areas without retail stores based on the daily sales collected from the ZIP codes with retail stores through the National Retail Data Monitor (NRDM). This simulation allows us to apply NRDM data in addition to other data sources in a multivariate analysis in order to rapidly detect outbreaks.

Introduction
In disease surveillance, an outbreak is often present in more than one data type. If each data type is analyzed separately rather than combined, the statistical power to detect an outbreak may suffer because no single data source captures all the individuals in the outbreak.[1] Researchers, thus, started to take multivariate approaches to syndromic surveillance. The data sources often analyzed include emergency department (ED) data, categorized by chief complaint; over-the-counter (OTC) pharmaceutical sales data collected by the National Retail Data Monitor (NRDM), and some other syndromic data.[1,2]

Methods
Owing to the limit of the existing dataset collected in NRDM, in that it does not have information about home ZIP codes of the patients for each of the product sales, we proposed a data simulation model to allocate the counts of OTC sales in patient's residential ZIP code areas.

To illustrate, we use an example of OTC medication purchases made by the patients living in six ZIP code areas with or without pharmacy stores (Figure 1). The nodes are connected by three types of arrows representing different types of commuting we presume: (1) for people who live in the ZIP code areas with pharmacy stores, they purchase OTC medications from those stores; (2) for people who live in ZIP code areas without stores, they will purchase OTC medications from (a) the adjacent ZIP code areas that have stores (solid arrows) (b) the nearest with-store ZIP code areas if neither their living ZIP codes nor the adjacent have stores (dashed arrows) or (c) their working ZIP code area with stores (doubled arrows).

Our methods consist of three steps. First, we split each non-store node into sub-nodes so that each sub-node only has one arrow going out. In the rightmost graph in Figure 1, s_{ij}^{w} represents the population of work flow between s_i and s_j, which was collected during the 2000 census, and s_{ij} represents the remaining population in area S_i who purchased OTC medication in area S_j, which is computed as proportional to the population of its target node. Second, for each with-store node, the sales counts are then re-allocated to all of its incoming nodes and itself assuming a multinomial

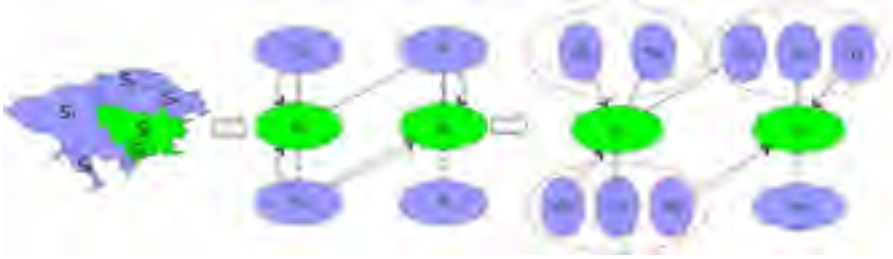

Figure 1 Modeling OTC medication purchases made by the patients living in six ZIP code areas. The leftmost figure shows their geographic relations. The green areas represent the ZIP codes with stores and the blue ones represent the ones without. The middle graph is to illustrate the three types of commuting in between. The right graph shows the sub-nodes after splitting.

Figure 2 The simulated OTC counts in Allegheny County.

distribution. Third, we combine the sub-nodes back into the original node by adding the allocated counts together.

Results

Figure 2 is an example of the simulated counts in Allegheny County, Pennsylvania. The model re-allocated the counts from 53 ZIP code areas with stores (in green) to the remaining 44 ZIP codes without stores.

Summary

We have presented a method to simulate the counts of purchased OTC medications in terms of residential location of the patients. This dataset can be used in multivariate analysis in combination with the syndromic dataset collected during the ED visits of patients in order to improve the power of early outbreak detection.

Acknowledgements

This paper was presented as an oral presentation at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA on 1–2 December 2010.

References

1 Kulldorff M, Mostashari F, Duczmal L, Katherine Yih W, Kleinman K, Platt R. Multivariate scan statistics for disease surveillance. *Stat Med* 2007;**26:**1824–33.
2 Wagner MM, Tsui FC, Espino J, Hogan W, Hutman J, Hersh J. National Retail Data Monitor for public health surveillance. *Morb Mortal Wkly Rep* 2004;**53** (Suppl): 40–2.