Evaluation of Over-the-Counter Pharmaceutical Sales As a Possible Early Warning Indicator of Human Disease

Steven F. Magruder

Surveillance of over-the-counter (OTC) pharmaceutical sales may someday be used to give public health officials an early warning about developing public health conditions. To acquire confidence in this surveillance, it is necessary to establish firmly that OTC sales may be used to predict changes in clinical conditions with a useful time lead. This article reports the magnitude of correlations between some OTC sales data and clinical data and measures the time lead after controlling for day-of-week effects and some holiday effects. Results, derived for several subregions of the National Capital Area, indicate about a 90% correlation between flu-remedy sales and physician diagnoses of acute respiratory conditions. The sales in question tend to occur approximately 3 days prior to the physician–patient encounters.

INTRODUCTION

Over-the-counter (OTC) pharmaceutical sales offer several advantages as a possible early indicator of public health. First, these products are widely used. According to a recent health survey,[1] 77% of the U.S. population reported having used nonprescription medications to treat a health condition at least once in a 6-month period. This compares to 43% who said they consulted a physician in the same time frame and 38% who said they used prescription medications. A second advantage of OTC sales is that a reliable and detailed electronic record is made instantly at the time of each sale, and such records are aggregated regionally for commercial purposes. These records may be readily transmitted to aid in health surveillance. Another possible advantage, which has not been so well established, is the timeliness of OTC sales data relative to other observable events that might occur when the public health is threatened.

APL receives OTC sales records routinely for analysis. The records include the daily sales volume of each of several hundred possible products and the store location of each purchase. (None of the records include purchaser identity information. Records provide numbers and types of products purchased only.) The total latency from time of sale until the data enter the surveillance system at APL ranges from 1 to 3 days, depending on the data source.

The purpose of this article is to present evidence that, for routine flu-like symptoms, OTC sales precede physician encounters by several days, on average.

OVERVIEW OF THE OTC DATA

APL is currently collecting large quantities of daily OTC sales data. We receive sales records of 622 different products under the general category of "cold remedies" from a single vendor, with similar numbers from other vendors. Many of the products are used to treat similar conditions. To analyze the data, we first use product groups that have been defined subjectively by a local expert in pharmacoepidemiology. The groupings are based on a product's presumed use and are further divided into child and adult medication. The product groups are summarized in Table 1. We aggregate the sales of individual products within each group to form a time series of daily sales of product packages. No attempt is made to apply different weights to different products, for example, by the total dosage contained in a package. Whether such a weighting scheme would be useful remains an open question.

Product sales from some of these product groups are known to be well correlated with the corresponding clinical data. (Clinical data again are anonymous and include information on numbers of patients diagnosed with a particular ailment but not the identity of the patients.) An example of this relationship is shown in Fig. 1, where chest rub product sales are compared with counts of physician diagnoses of acute bronchitis
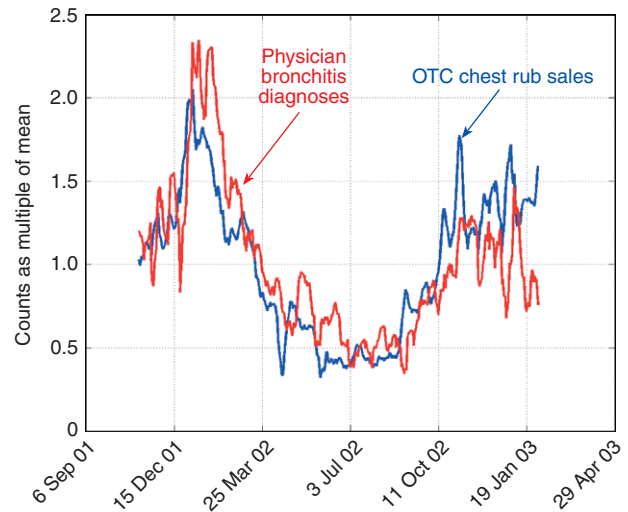

Figure 1. Over-the-counter (OTC) sales of chest rub products compared with physician diagnoses of bronchitis or bronchiolitis in the National Capital Area. The dates for the OTC data indicate date of sale; for diagnoses, they indicate the date on which patients presented themselves to physicians.

or acute bronchiolitis. The diagnosis data are based on insurance claims, which are coded by the physician according to the International Classification of Diseases, Ninth Revision (ICD-9) standards (see, e.g., Ref. 2). The date associated with these claims is the date of presentation of the patient. The date associated with the OTC data is the date of sale. Both time series were smoothed using a 7-day moving average to improve comparisons between the two data sources. This was done because the physician claim data exhibit very strong day-of-week dependence, while the OTC data exhibit a weaker and different day-of-week dependence.

To obtain the curves in Fig. 1, we divide each time series by its respective mean value over the time period of the plot. Both types of data are drawn from the National Capital Area (NCA), which includes Baltimore, Washington, and the Baltimore–Washington Corridor. It is interesting to note both the high degree of correlation between the two data types and the time lead of the OTC data relative to the clinical data. The cross-correlation between the two curves is shown as a function of time lag in Fig. 2. We found a peak correlation value of 0.86 for a 7-day lag of bronchitis diagnoses relative to chest rub sales.

Another example of strong correspondence between OTC and clinical data is shown in Fig. 3, where OTC flu remedy sales are compared with physician diagnoses of a variety of acute respiratory conditions that are currently monitored by the ESSENCE (Electronic Surveillance System for the Early Notification of Community-based Epidemics) surveillance system. (This ESSENCE Respiratory Group includes 169 separate ICD-9 codes. A few of the most common are "Acute Upper Respiratory Infections of Unspecified Site," "Acute Pharyngitis,"
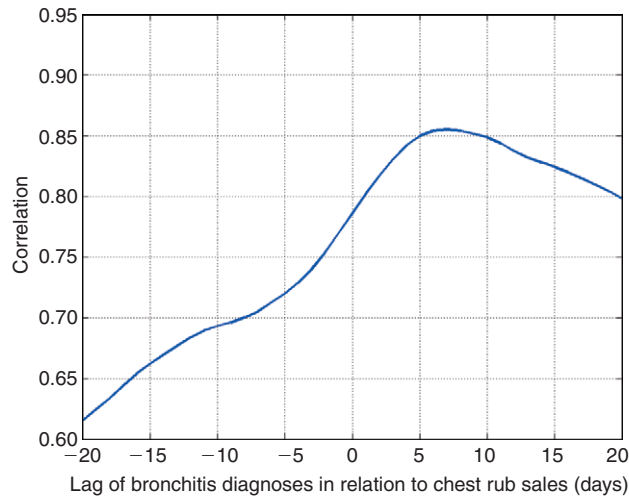
Table 1. Over-the-counter medication product groups.

Product group name	Number of products
Adults	
Allergy	6
Chest rub	1
Cold	11
Cold drops	10
Flu	3
Nasal spray	6
Syrup	29
Throat spray	2
Other	8
Children	
Allergy	62
Chest rub	9
Cold	204
Flu	36
Inhaler	9
Lozenge	4
Nasal spray	59
Syrup	121
Throat spray	11
Other	30

Figure 2. Cross-correlation versus lag time for the two curves shown in Fig. 1.

"Acute Bronchitis," "Unspecified Viral Infection in Conditions Classified Elsewhere and of Unspecified Site," and "Cough," but there are many important contributors. See the article by Lombardo elsewhere in this issue.) The data in Fig. 3 were processed in the same way as the data in Fig. 1. In this case, the maximum correlation is 0.89, observed when the data are adjusted for a 12-day lead of the OTC data.

MEASUREMENT OF TIME LEAD

The perceived value of OTC sales as a data source for syndromic surveillance would be greatly enhanced if it could be shown that such sales provide an earlier indicator than clinical data for health problems. The time lead can be estimated by simply correlating one time series against the other and looking for a peak, as discussed above. However, the simple time lead measurements of
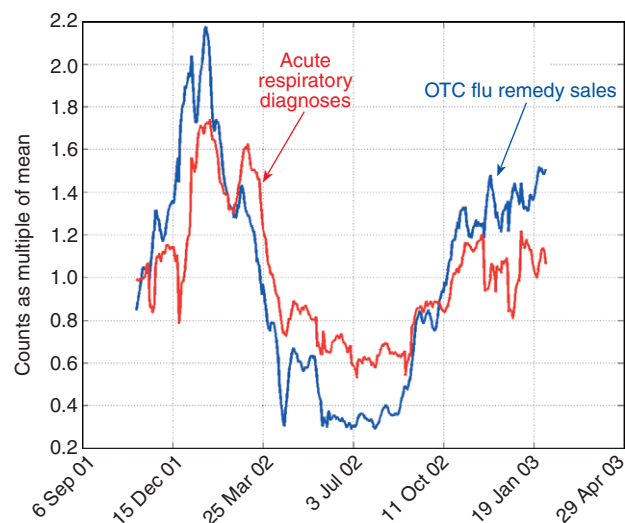

Figure 3. Over-the-counter sales of flu remedies compared with physician diagnoses of acute respiratory conditions in the National Capital Area.

the previous section cannot be considered to be completely reliable for reasons presented below.

In addition to day-of-week effects, the data are also subject to holiday effects. The weeklong Christmas–New Year holiday effect is obvious in the physician data in Figs. 1 and 3: the physician data drop dramatically during that week, while the OTC data do not. The correlations are reduced, and the comparison between the two data sources is obscured by mixing the holiday behavior with the non-holiday behavior. In our present analysis, therefore, we censor the major holiday period, starting 3 days before Christmas and ending 3 days after New Year's Day. (The 3-day buffer is added to account for the effects of the 7-day moving average.) Data for the censored period are not included in our calculations. Other holidays, notably Thanksgiving, also have measurable effects (e.g., in Figs. 1 and 3), but they represent shorter time periods, and the effects are less dramatic. For this analysis, we do not censor those periods.

Another limitation is the number of independent samples of time lead. Although the data series in Figs. 1 and 3 represent hundreds of days of observations, and each observation typically involves hundreds of transactions, the measured time lead is sensitive to the seasonal rise and fall of the respective curves. The factors affecting seasonal trends are not completely understood, but they could be influenced by local weather patterns, the timing and virulence of influenza epidemics (which in turn could be influenced by global weather or random genetic mutations), the status of the economy, characteristics of the local population, or other unknown conditions. The point is that some trends in the data may be systematic for one location and season, but vary from season to season or from location to location. Therefore, a measurement of time lead is most meaningful if it can be made for multiple locations and over multiple seasons.

Figures 1 and 3 clearly show great variation between the 2001–2002 winter and the 2002–2003 winter. Year-to-year variability is not addressed here; the current time lead analysis is confined to the fall 2001–spring 2002 period. To address the potential spatial variability, we aggregated data separately within six distinct spatial regions (Fig. 4). The resulting OTC sales and physician diagnoses time series are shown in Fig. 5. The time period within the red vertical bars is excluded from correlation analysis owing to holiday effects. In each plot, the blue curve represents sales of OTC flu remedies and the green curve represents physician diagnoses of respiratory conditions based on ESSENCE codes.

Although the general characteristics of the 2001–2002 winter cold season were similar in the six regions (Fig. 4), significant variability is also apparent. For example, the peak activity in both physician claims and OTC sales for the Eastern Shore region occurs approximately 3 weeks later than in the Urban NCA.
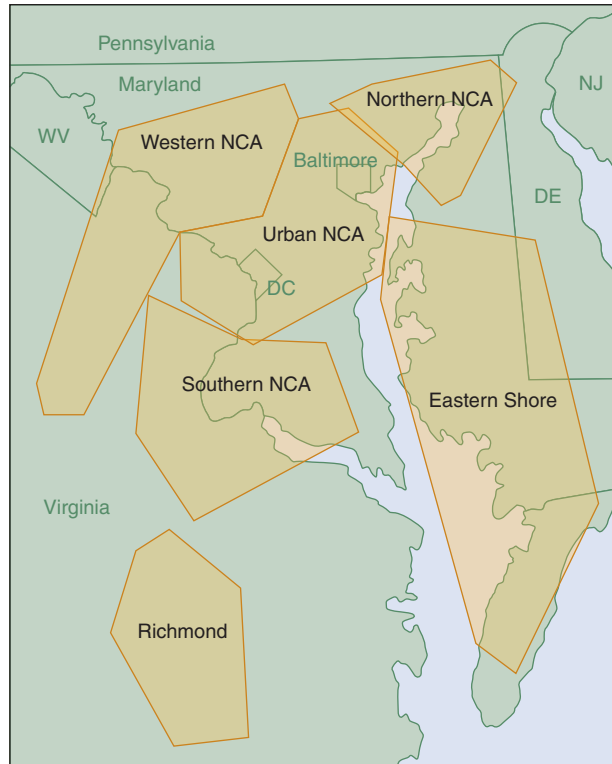
Figure 4. Map of the National Capital Area (NCA) showing the spatial areas used in the correlation analysis.

The correlations and peak correlation leads of the OTC sales relative to physician visits is summarized in Table 2. The mean of these time leads is 2.8 days. If one assumes that these measured leads are independent draws from a single unknown normal (or Gaussian) distribution, then the hypothesis that the OTC sales are, in fact, lagging the physician visits can be rejected with 93% confidence, given these results.

CONCLUSIONS

The results presented in this article indicate that sales of OTC flu remedies were well correlated with physician diagnoses of acute respiratory conditions throughout the NCA (and slightly beyond) in the 2001–2002 winter cold season. The results also indicate a consistent time lead of flu remedy sales, which is estimated to be approximately 3 days. Chest rub product sales also show a strong correspondence with bronchitis diagnoses. These results tend to

strengthen the hypothesis that OTC product sales can be used as an early indicator of human disease in a particular geographic area.

One remaining question is whether the observed time lead is related to habitual buying behavior in every cold season or to the time span between self-medication and professional intervention, given the actual occurrence of illness. If it is the former, the time lead is not necessarily useful for population health surveillance. If it is the latter, the time lead could be very useful. One way to address this issue is to control both data sources for their seasonal behavior and look for correlations between the short-time deviations from the seasonal behavior. For the diagnosis data, these short-term deviations are attributable to health-related events, and their exact timing is not predictable by the public, so correlations and time leads in these season-controlled residuals will provide some additional evidence as to the cause of the correlations. Such an analysis is currently under way at APL.

Another major area of research relates to the formation of product groups in the OTC data and syndrome groups in the clinical data. For geographic regions of the size considered in this study, individual product sales and individual ICD-9 diagnoses often do not provide sufficient sample sizes for statistically meaningful conclusions. Therefore, grouping of these products and codes is essential, and it is also essential to understand which product groups best predict which diagnosis groups. Research into these questions is also in progress.
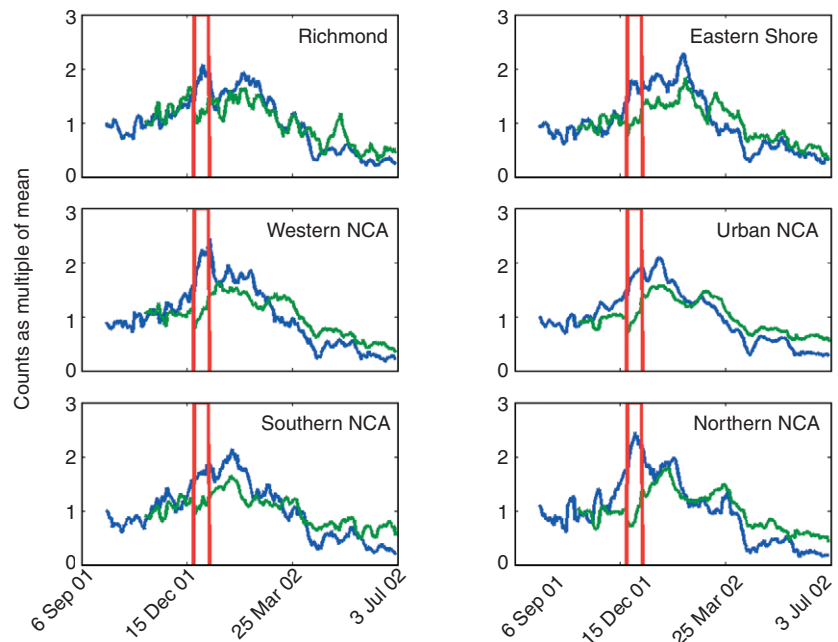

Figure 5. Over-the-counter sales of flu remedies (blue curves) compared to physician diagnoses of acute respiratory conditions (green curves) in the six regions defined in Fig. 4. The time period within the red vertical bars is excluded owing to holiday effects.

Table 2. Correlation analysis results for six regions in or near the National Capital Area.

Region	Mean weekly OTC count	Mean weekly physician count	Peak correlation	Lead time of OTC sales (days)
Urban NCA	1840	2823	0.93	5
Southern NCA	104	225	0.91	−2
Western NCA	119	368	0.90	7
Richmond	166	61	0.90	0
Eastern Shore	82	137	0.87	2
Northern NCA	97	499	0.86	5

REFERENCES

[1]*Self-Care in the New Millennium*, Roper Starch Worldwide, Inc., prepared for the Consumer Healthcare Products Association (2001), available at http://www.chpa-info.org/statistics/Roper-Starch.asp (accessed 7 Aug 2003).
[2]*Diagnostic Coding Essentials*, Ingenix Publishing Group, Salt Lake City, UT (2001).

THE AUTHOR

STEVEN F. MAGRUDER is a member of the APL Principal Professional Staff. He received a Ph.D. in physics in 1978 from the University of Illinois at Urbana. He joined APL that same year, and has spent most of the past 25 years developing physical modeling and detection/signal processing algorithms for a variety of submarine sensing technologies. Dr. Magruder is currently developing tools for the early recognition of a biological attack and focusing on the evaluation of data sources and algorithms. His e-mail address is steve.magruder@jhuapl.edu.