BMC Medical Informatics and Decision Making

BioMed Central

Research article

Open Access

Estimation of hospital emergency room data using otc pharmaceutical sales and least mean square filters

AH Najmi* and SF Magruder

Address: National Securit Technology Department, The Johns Hopkins University, Applied Physics Laboratory Laurel, MD 20723-6099 U.S.A

Email: AH Najmi* - najmi@jhuapl.edu; SF Magruder - steve.magruder@jhuapl.edu

* Corresponding author

Published: 15 March 2004

BMC Medical Informatics and Decision Making 2004, **4**:5

This article is available from: http://www.biomedcentral.com/1472-6947/4/5

Received: 30 October 2003
Accepted: 15 March 2004

Abstract

Background: Surveillance of Over-the-Counter pharmaceutical (OTC) sales as a potential early indicator of developing public health conditions, in particular in cases of interest to Bioterrorism, has been suggested in the literature. The data streams of interest are quite non-stationary and we address this problem from the viewpoint of linear adaptive filter theory: the clinical data is the primary channel which is to be estimated from the OTC data that form the reference channels.

Method: The OTC data are grouped into a few categories and we estimate the clinical data using each individual category, as well as using a multichannel filter that encompasses all the OTC categories. The estimation (in the least mean square sense) is performed using an FIR (Finite Impulse Response) filter and the normalized LMS algorithm.

Results: We show all estimation results and present a table of effectiveness of each OTC category, as well as the effectiveness of the combined filtering operation. Individual group results clearly show the effectiveness of each particular group in estimating the clinical hospital data and serve as a guide as to which groups have sustained correlations with the clinical data.

Conclusion: Our results indicate that Multichannle adaptive FIR least squares filtering is a viable means of estimating public health conditions from OTC sales, and provide quantitative measures of time dependent correlations between the clinical data and the OTC data channels.

Background

Surveillance of Over-the-Counter (OTC) pharmaceutical sales as a potential early indicator of developing public health conditions has been suggested in the literature [1]. OTC sales offer several advantages as possible early indicators of public health. They are, first of all, very widely used. According to a recent health survey [2], 77% of the U. S. population said they have used non-prescription 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 period, and 38% who said they used prescription medications.

A second advantage of OTCs is that reliable and detailed electronic records are made at the time of sale. These records are aggregated regionally for commercial purposes. The only additional burden for health surveillance purposes is to communicate this data to the appropriate public health organizations. The OTC data contain significant information, e.g. sales volume of each of several hundred possible products, and the precise location of the store where they are sold.

A third possible advantage, which has not been so well established, is the timeliness of OTC sales relative to other

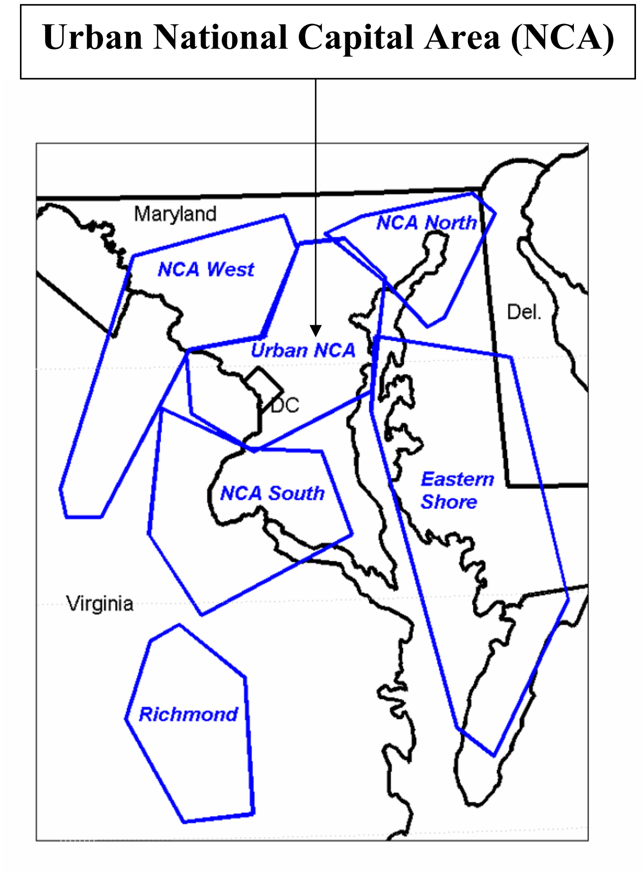

Figure 1
Map of the Urban National Capital Area (NCA) in the United States.

observable events that might occur when the public health is threatened.

The purpose of this article is to present evidence that when judiciously grouped, the OTC data show time-dependent correlations with clinical data, and that the latter can be reconstructed from the former using a linear filter.

JHU/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 these products are used to treat very similar conditions. As a starting point to analyze the data, we made use of product groups that had been defined subjectively by a local expert in the domain of 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 (see Additional file 1). We aggregated the sales of individual products within each group to form a time series of daily sales of product packages. No attempt was 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. These product groups are further divided into children's medications and adult medication.

Product sales from some of these product groups are known to be good indicators of the corresponding clinical data [3]. For instance, chest rub sales are highly correlated with the count of physician diagnosis of acute bronchitis or acute bronchiolitis [4].

The hospital data consisted of daily counts of Emergency Room visits that were assigned discharge diagnoses from a list of acute respiratory conditions. The detailed list is quite long and is available from the authors upon request and is not tabulated in this paper. This list was chosen because of its use at the time in the ESSENCE surveillance system [5]. All visits to Maryland emergency rooms were included for patients living in the Urban_National Capital Area region, and with ages 18 and above. The data were binned according to the day of encounter. The Urban NCA includes (approximately) Baltimore, Washington, and the "Corridor" in between. A map of this region is shown in figure 1. This is basically the same geographic population represented in the OTC data streams.

Methods

Least mean square (LMS) filtering

Here, we consider the clinical data as the primary channel of an adaptive LMS filter. The OTC product groups are then used to estimate the daily clinical data in the following manner. Today's and several past days' OTC data are used together to find an estimate of today's clinical data, which is then compared to the actual value of today's clinical data to update the filter coefficients in such a way as to minimize the mean square error between the today's

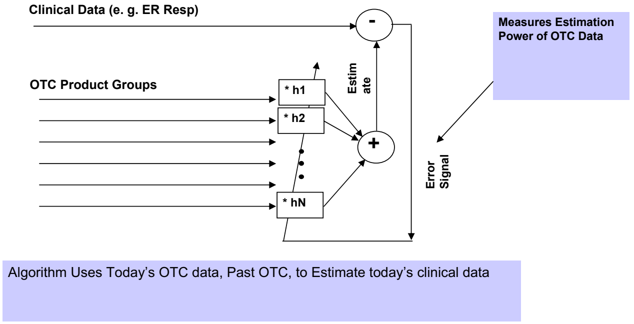
Clinical Data (e. g. ER Resp)
-
Measures Estimation Power of OTC Data
OTC Product Groups
* h1
* h2
* hN
Estim ate
+
Error Signal
Algorithm Uses Today's OTC data, Past OTC, to Estimate today's clinical data

Figure 2
Diagram of the multi-channel adaptive LMS filter

estimated and actual clinical data (see figure 2). This method addresses the pure filtering problem, in the parlance of optimal filter theory, and we now describe the problem in mathematical terms.

If we denote the clinical data time series by $y[n]$, and the OTC reference time series channels $x_j[n]$, where the index n denotes the day number and the index j denotes the OTC product group, then the today's and past days' OTC data are used to estimate today's clinical data, in the sense that the estimated quantity is $\hat{y}[n] = \sum_{j=1}^{J} \sum_{m=0}^{M-1} x_j[n-m]h_j[m]$, and it is to be compared directly to the actual value of the clinical data today. This is referred to as the "filtering" problem, as distinct from the following two problems [6]. The "prediction" problem attempts to estimate future values of the clinical data using today's and past days' values of the OTC channels, i.e. the predicted quantity is $\hat{y}[n+k] = \sum_{j} \sum_{m} x_j[n-m]h_j[m]$, $k > 0$, which is then to be compared to the actual value of the clinical data on day number $n + k$. The "smoothing" problem, of no interest to us in this application, is to compute past values of the clinical data using today's and past values of the OTC data, and so the "smoothed" value is $\hat{y}[n-k] = \sum_{j} \sum_{m} x_j[n-m]h_j[m]$, $k > 0$, which is to be compared to the actual value of the clinical data on day number n - k.

It should be clear from the above description that a useful public health surveillance system could be quite interested in the "prediction" problem. Clearly, cross correlations between the OTC data channels, if they exist in sufficient amounts, could be used to "predict" the clinical data from OTC channels. Similarly, auto correlations of

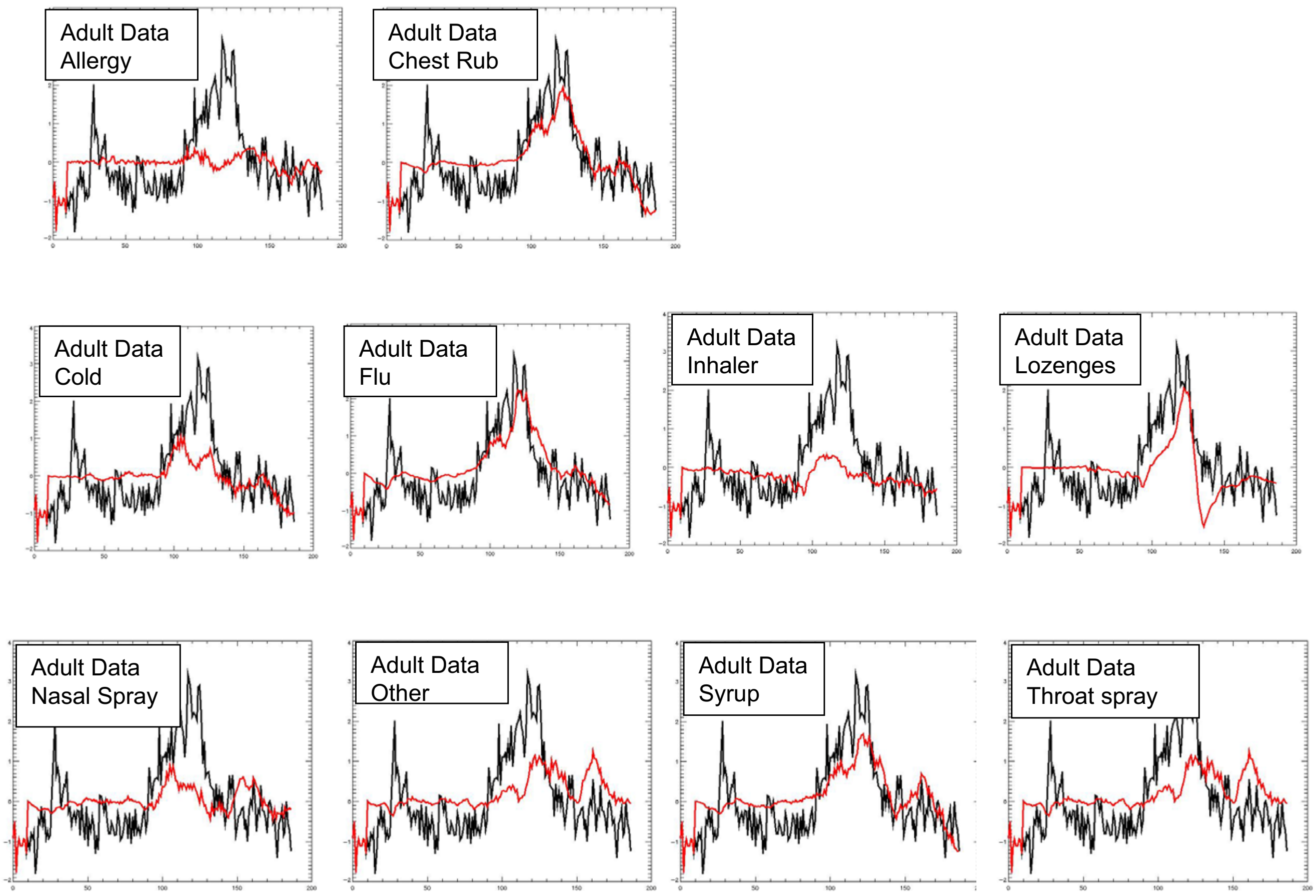

Figure 3
Individual LMS estimation outputs together with actual hospital data.

the clinical data itself, if they exist in sufficient amounts, could be used to (self) "predict" the clinical data. If both types of predictions can be made reasonably successfully, then one could compare them and/or combine them in order to maximize the "prediction" abilities of some component of a useful surveillance system. This latter problem is particularly difficult and in order to motivate a serious attempt at it, we have set ourselves the simpler task of studying the "filtering" problem. This way we can use the adaptive filter technique to show "time dependent" correlations between the clinical and OTC channels, that are caused by the non stationary nature of both data sets.

The requirement of minimum mean-squared error can be most easily implemented using the Widrow LMS algorithm [6]. A related adaptive filter that we have used in this paper is the normalized LMS algorithm that avoids some of the difficulties with the choice with the adaptation parameter. This is the solution to the following constrained optimization problem. Given the primary channel data $p[n]$ and the reference channels' data $\underline{z}_j[n]$ the optimal filter coefficients $\underline{h}_j[n + 1]$ are found by minimizing the magnitude of the difference $||\underline{h}[n + 1] - \underline{h}[n]||$ subject to the constraint $p[n] = \underline{h}_j^T[n+1]\underline{z}_j[n]$. We show in the Appendix (see Additional file 2) that the adaptive filter must satisfy the following equations:

$$\hat{p}[n] = \sum_{j=1}^{N} \sum_{m=0}^{M-1} h_n^j[m] z_j[n-m]$$

$$e[n] = p[n] - \hat{p}[n]$$

$$h_{n+1}^j[m] = h_n^j[m] + \frac{\mu}{a + \left\| \underline{z}[n] \right\|^2} e[n] z_j[n-m]$$

where $0 < \mu < 2$ and $0 < a << 1$.

Results and discussion

Figure 3 shows the estimation outputs for each OTC product group, and figure 4 shows the output of a multichannel filter (all products included) superimposed on the actual Emergency room data for the same period of time, for adults. Based on these figures and defining estimation power $1 - \frac{\sigma^2_{\text{residual}}}{\sigma^2_{\text{original}}}$, we show in figure 3 the effectiveness of each product group in estimating the Emergency room data. Note that $\sigma^2_{\text{residual}}$ is the quantity minimized by the LMS filter adaptation equations. We have normalized it by the variance of the input clinical data (the "primary" channel), and subtracted it from 1, so that numbers closer to 1 (from below) imply the best estimation results.

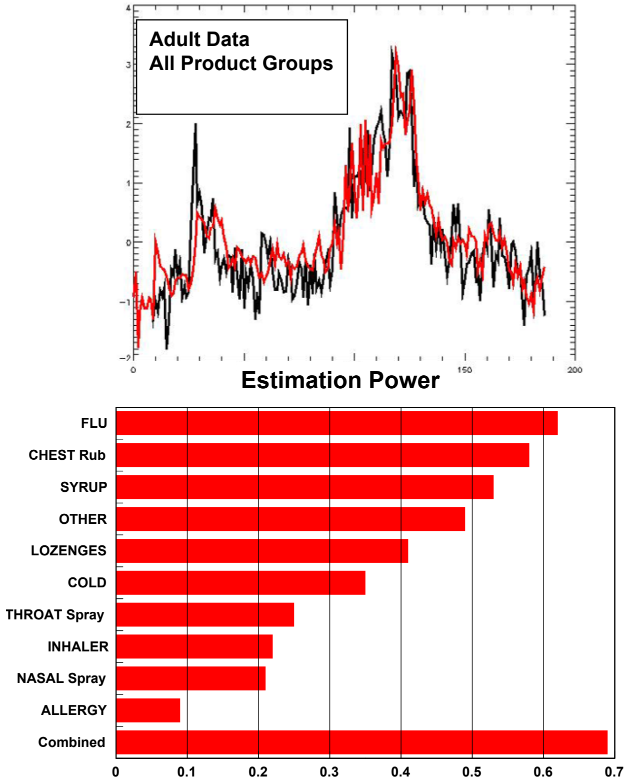

Figure 4
Multiple channel LMS estimation output together with actual hospital data, and graph of estimation power of individual LMS filters as well as the Multi-channel one

As is apparent from the figure 4 we have been quite successful at reconstructing the emergency room data from all 10 product groups, except the large peak at day 20 that has not been estimated well. This clearly shows that the latter event is not correlated with any of the OTC data streams at the same time. Individual group results clearly show the effectiveness of each particular group in estimating the hospital data and serve as a guide as to which groups are more likely to produce better prediction of the hospital data, in the sense described in the Methods section above.

Conclusions

The perceived value of OTC sales as a data source for syndromic surveillance would be greatly enhanced if it could be shown that OTC sales provided an earlier indicator for health problems than could be obtained from clinical data. The results presented in this article indicate that sales of over-the-counter flu remedies were well correlated with physician diagnoses of acute respiratory conditions throughout the National Capital Area (and slightly beyond) in the 2001–2002 winter cold season and can reproduce that data with rather small error. These results tend to strengthen the hypothesis that some OTC product sales might be used as an early indicator of a general class of human disease known as acute respiratory condition, if we are successful in extending the estimation algorithms presented here to one for predicting the same results with non zero positive time lags.

Competing interests

None declared.

Abbreviations

otc: over the counter (medications).

Additional material

Additional File 1

Table 1 - OTC Adult Medication Product Groups
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6947-4-5-S1.pdf]

Additional File 2

Appendix - The normalized LMS algorithm derivation.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6947-4-5-S2.pdf]

Acknowledgements

This research is sponsored by the Defense Advanced Research Projects Agency and managed under Naval Sea Systems Command (NAVSEA) contract N00024-98-D-8124. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, NAVSEA, or the United States Government.

We would like to thank Sheri Happel-Lewis who provided the OTC cold-product subgroupings used in this study.

References

1. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE: **Early statistical detection of Anthrax outbreaks by tracking over-the-counter medication sales.** *PNAS* 2002, **99:**5237-5240.
2. **Self-care in the New Millennium, report by Roper Starch Worldwide, Inc., prepared for the Consumer Healthcare Products Association, Roper-Starch.** 2001.
3. Magruder S: **Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of public health.** *Johns Hopkins University Applied Physics Laboratory Technical Digest* 2003, **24:**. (to appear)
4. **Diagnostic Coding Essentials.** *Ingenix Publishing Group, Salt Lake City, Utah*; 2001.
5. Lombardo J, Burkom H, Elbert E, Magruder S, Happel Lewis S, Loschen W, Sari J, Sniegoski C, Wojcik R, Pavlin J: **A Systems Overview of the Electronic Surveillance System for the Early Notification of Community Based Epidemics (ESSENCE II).** *Journal of Urban Health* 2003, **80:**i32-i42.
6. Moon T, Stirling W: **Mathematical methods and Algorithms for Signal Processing.** *Prentice Hall*; 2000.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6947/4/5/prepub