Reviewer's report

Title: Estimation of Hospital Emergency Room Data Using OTC Pharmaceutical Sales and Least Mean Square Filters

Version: 1 Date: 30 November 2003

Reviewer: David Buckeridge

Reviewer's report:

General
******

The authors present an approach to forecasting the number of visits to emergency departments (ED) for acute respiratory conditions. They use a least mean-square filter of historical ED visits and historical and current over the counter pharmaceutical (OTC) sales to predict current ED visits. Their stated goal is to demonstrate time-dependant correlations between the OTC and ED data streams.

The accurate forecasting of clinical visit frequency is an important problem in public health surveillance. OTC data may be of use for detection of disease outbreaks, and a deeper understanding of the properties of these data for surveillance is of interest. The authors are in effect simultaneously examining a novel forecasting method (the least mean-square filter) and a novel surveillance data source (OTC sales). This makes it difficult to understand the relative contribution of the forecasting method (in comparison to standard methods) and the data sources (historical ED visits, historical and present OTC sales) to prediction. In addition, the authors do not clearly address their goal of demonstrating the time-dependence of the correlations between the ED and OTC data streams. While the relative estimation power of various OTC product groups is presented, the estimation power of historical ED visits is not presented, and the authors do not present any information on the time-dependence of the correlation between ED visits and OTC sales.

Discretionary Revisions (which the author can choose to ignore)
*************************************************************************

Page 2,3 – The authors could describe the least mean-square algorithm more clearly. For example, I assume that ‘n’ refers to the current time interval, but this is not explicitly stated.

Page 3 – The first paragraph of the Results and Discussion section is something that one would usually expect to find in the Methods section. Also, my understanding is that the paper cited does not provide the actual coding categories for the syndromes. In the spirit of reproducibility, it would be helpful to provide the diagnostic codes used in the study as an Appendix.

Page 3 – There is little consideration of the potential role of the least mean-square algorithm in surveillance. A discussion of this method in relation to more traditional approaches such as time-series and wavelets would be helpful.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
*************************************************************************

Page 1 – The authors suggest that OTC sales data offer an advantage over physician visit data due
to the greater frequency of OTC use. This is somewhat misleading as they present data on OTC sales, not OTC use. Each physician visit generally results in an electronic record, while there are likely to be many OTC uses to each OTC sale, which results in an electronic record.

Page 2 – The authors state that product sales in some groups are known to be good indicators of clinical data. A reference should be provided to support this assertion.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Page 2 – The authors present the titles of the OTC product groups, but no information on the process for arriving at these groupings, or the actual products falling into each groups. Both the process for creating the groups and the actual groupings should be shown as the categorization may have considerable influence on the results. As space is not an issue, the grouping could be presented as an Appendix.

Page 2 – The authors should give a better description of the data sources, and identify their relationship. Presumably the data sources reflect the same population to some extent (i.e., same geographic region, similar individuals likely to appear in both data sources). The extent of this overlap should be explicitly described to facilitate interpretation of the results.

Page 2,3 – The authors do not describe how the number of prior values for a reference channel is selected (i.e., the selection of the M for a given channel j). Presumably the choice of the maximal lag is related to the frequency of temporal patterns that one is attempting to incorporate into the forecast (e.g., day of the week, season). It is important to understand the influence of this choice on the performance of the filter.

Page 3 – The definition of estimation power is not clear and the authors do not identify why they choose this metric as opposed to something such as mean-square error. The authors should expand and clarify the definition and provide a citation for their chosen metric.

Page 3 – There is no explicit consideration of the temporal correlation between the ED and OTC data at different temporal lags. It would be helpful if the authors could present some information on this topic. For example, might it be possible to identify the contribution of OTC data to ED prediction at different temporal lags by examining the average OTC filter coefficients at different temporal lags (i.e., look for the maximum average h_j[m] across different m for the each j)? Also the authors identify that historical ED data is one of the reference channels, but no specific results are presented for the estimation power of this channel. Is this channel included in the chart shown in Figure 3?

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests: None