Data Normalization in Biosurveillance: An Information-Theoretic Approach

William Peter, PhD, Amir H. Najmi, D.Phil., Howard Burkom, PhD
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

Abstract
An approach to identifying public health threats by characterizing syndromic surveillance data in terms of its surprisability is discussed. Surprisability in our model is measured by assigning a probability distribution to a time series, and then calculating its entropy, leading to a straightforward designation of an alert. Initial application of our method is to investigate the applicability of using suitably-normalized syndromic counts (i.e., proportions) to improve early event detection.

Introduction
In this work we discuss an approach to identifying public health threats and disease outbreaks by characterizing syndromic surveillance data in terms of its information content, or surprisability. The metric for information content will be defined by the Shannon entropy\(^1\) \(H(X)\) of a discrete random variable \(X\) defined over the sample space of the values \(x_i\) of the time series.

Given two time series \(\{x(t)\}\) and \(\{y(t)\}\), their mutual information \(I(X,Y)\) is the average number of bits of \(\{x(t)\}\) that can be predicted by measuring \(\{y(t)\}\). It is also interpreted as the difference between the uncertainty of \(\{x(t)\}\) and the remaining uncertainty of \(\{x(t)\}\) after observing \(\{y(t)\}\).

Mutual information is a generalization of linear correlation since it is defined in terms of the entropy, which uses the full probability density function (and not the low-order moments) of the random variable. It can thus measure dependence on data that vary nonlinearly.

Investigating Alerts by Counts or Proportions
Applications of automated disease surveillance are often limited to opportunistic data streams in which a true denominator for traditional incidence rate calculations is unavailable. These systems monitor either daily syndromic diagnosis counts themselves or the ratio of such counts to an unfiltered sum such as all daily records in the database. Figure 1 depicts a numerator (blue) and denominator (green) series for such a ratio. The advantage of monitoring the ratio is that the denominator series contains information that can explain away features of the numerator series that are irrelevant to health surveillance, such as cyclic or seasonal patterns and changes in population behavior or in the participation of data providers.

Our approach is to use information-theoretic techniques to determine under what conditions proportions are preferable to counts for separating potential public health signals from background noise in monitor data streams. The approach should also readily identify effective numerator-denominator series combinations. Finally, it may be possible to use these techniques to develop novel detectors for public health threats.

Figure 1. Example of daily syndrome counts (blue) and total diagnostic counts (green) used in our study.

Conclusion
 Appropriately conditioning and combining syndromic data streams is important to achieve the signal-to-noise gains necessary for increased sensitivity in biosurveillance. Information-theoretic methods, little used in disease surveillance up to now, will be applied to gain practical guidance toward this goal,

References