Statistical analysis algorithm for point of care surveillance in refugee/migrant reception centres in Greece

Purpose

The purpose of the proposed algorithm is to calculate the **expected proportional morbidity** for each syndrome under surveillance, both in total and per reception centre. It also serves to calculate an **alert level** that objectively separates normal morbidity levels from unexpected outbreaks and epidemics.

The algorithm is based on the method of Farrington et al[1], which is simple, flexible and widely used for surveillance and outbreak detection purposes[2].

Input data

- *Y(t) :* number of cases of the syndrome under surveillance at time *t* (day or week of notification).
- *N(t) :* total number of patient visits at time *t* (day or week of notification).

Step 1: Fit a regression model

A quasi-Poisson regression model with overdispersion is fit on the above input data. From the model we calculate the expected number of cases *μ(t)* for a given total number of visits.

In the simplest case we use an intercept-only model, hence the expected proportional morbidity is stable over time, and equal to exp*(α)*:

$$\log(\mu(t)) = \alpha + \log(N(t)) \quad (1)$$

Further covariates *f(X)* can be added in this basic model:

$$\log(\mu(t)) = \alpha + \beta * f(X) + \log(N(t))$$

In this fashion, the expected proportional morbidity can be expressed as a function of time *t* or other factors, such as season, day of the week, temperature, fasting periods (e.g. Ramadan) etc.

In order to monitor long-term normal trends in morbidity we suggest adding time *t* in the model via a natural cubic spline function *S(t, k)*. If time *t* is expressed in days, we suggest setting the degrees of freedom *k* equal to the number of completed months in the time series. In practice this means that the time trend of the expected proportional morbidity is free to vary approximately on a monthly basis. Therefore we avoid short term fluctuations that would amount to an overfitted model.

$$\log(\mu(t)) = \alpha + \beta * S(t, k) + \log(N(t)) \quad (2)$$

For rare syndromes it is more reasonable to consider the expected morbidity as time invariant, unless the time series is several years long. Similarly, it is meaningless to incorporate time in the model if the time series is very short.

Consequently we suggest using model (2), unless:

- the time series is less than one month long, and/or
- the number of cases *Y(t)* is zero in at least 75% of observations

In this case model (1) is preferred.

Step 2: Calculate standard deviation and the alert level

In order to calculate the standard deviation for the expected number of cases *Y(t)* a 2/3-power transformation is applied[1], which corrects skewness characteristic of the Poisson distribution. The alert level $Y_\alpha(t)$ for the observed number of cases is finally calculated in terms of the number of standard deviations *Z* as follows:

$$(3) \quad Y_\alpha(t) = \mu(t) * \left(1 + \frac{2}{3} Z \sqrt{\frac{\varphi + Var(\mu(t))/\mu(t)}{\mu(t)}}\right)^{3/2}$$

where φ is the overdispersion factor calculated from the regression model (1) or (2). In similar fashion a z-score for the observed number of cases can be calculated, i.e. the number of standard deviations in relation to the expected number of cases.

For Z=1.64 equation (3) defines a one-sided 95% prediction interval for the observed number of cases *Y(t)*. For Z=2 standard deviations, the prediction interval is 97.5%, i.e. only 2.5% of observations is expected to fall randomly over the value of $Y_\alpha(t)$. **We propose setting 2 standard deviations as the alert level** (Z=2). Simply dividing $Y_\alpha(t)/N(t)$ produces the alert level for the observed proportional mortality.

Step 3: Discard unusually high values and recalculate

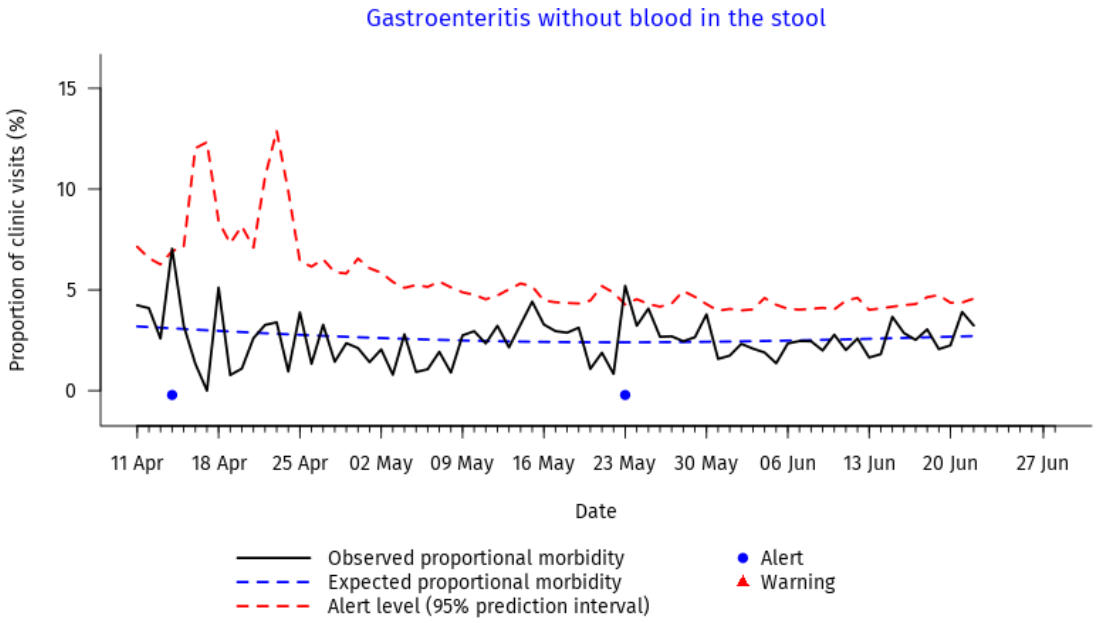
The regression model on step 1 estimates the expected number of cases corresponding to the usual morbidity patterns. Outbreaks and epidemics will increase the standard deviation of the expected number of cases, and raise the alert level. Therefore any observed values that are reliably part of an epidemic are best omitted from the calculation of expected morbidity and alert level. On the other hand, excluding values that are randomly high but not part of an epidemic will underestimate the standard deviation and inappropriately reduce the alert level.

As a result, we propose the following procedure: after steps 1 and 2, the time series is checked for any points with a z-score over 3, which almost certainly correspond to an outbreak. If such points exist, the regression model (step 1) is fit again omitting these points, and the alert level is calculated again (step 2). The cycle is repeated if nessesary, until all the points used in the model have a z-score less than 3.

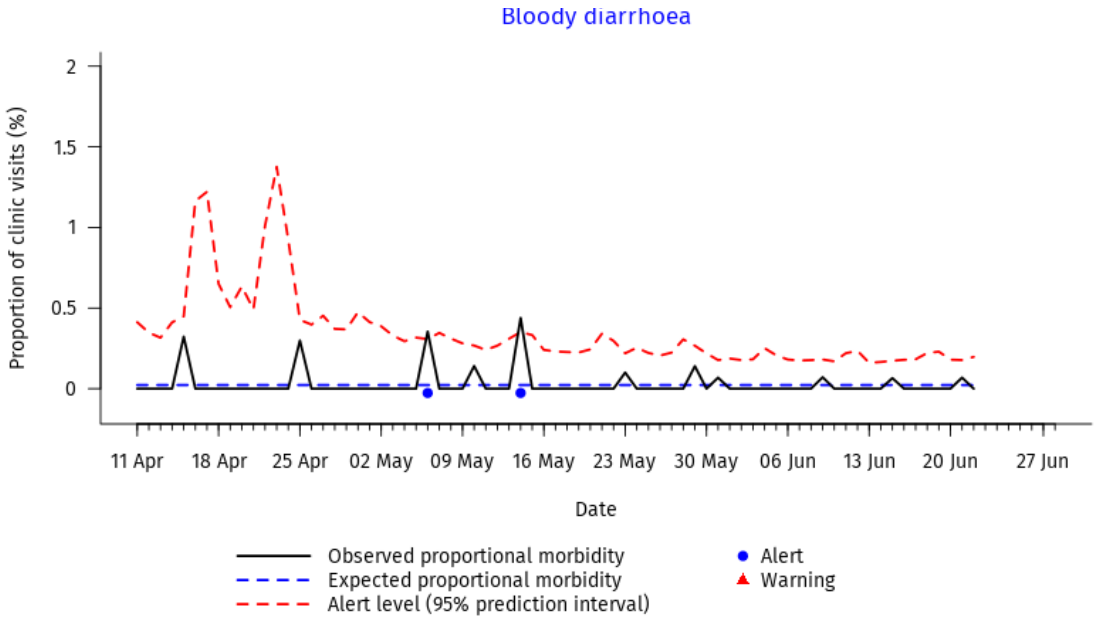
As a complementary procedure, it is possible to manually exclude values from the algorithm if their z-score is less than 3 but they are still considered part of a confirmed epidemic.

Example

On the first plot, the model includes a natural cubic spline with 2 degrees of freedom. Fluctuations of the alert level line are due to differences in the daily total number of patient visits (i.e. the denominator); if the number is lower, the alert level is higher.



On the second plot, the syndrome under surveillance is rare and an intercept-only model is used. The expected morbidity is stable over time and at a very low level. Fluctuations in the observed morbidity are caused by single cases, except at two occasions where two cases were observed in the same day; on both these occasions, the alert level was slightly exceeded.



Βιβλιογραφία

1. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. Journal of the Royal Statistical Society Series A (Statistics in Society). 1996;159(3):547.

2. Hulth A, Andrews N, Ethelberg S, Dreesman J, Faensen D, van Pelt W, et al. Practical usage of computer-supported outbreak detection in five European countries. Euro Surveill. 2010 Sep 9;15(36).