An improved algorithm for outbreak detection in multiple surveillance systems‡

Angela Noufaily,a*† Doyo G. Enki,a Paddy Farrington,a Paul Garthwaite,a Nick Andrewsb and André Charlettb

In England and Wales, a large-scale multiple statistical surveillance system for infectious disease outbreaks has been in operation for nearly two decades. This system uses a robust quasi-Poisson regression algorithm to identify aberrances in weekly counts of isolates reported to the Health Protection Agency. In this paper, we review the performance of the system with a view to reducing the number of false reports, while retaining good power to detect genuine outbreaks. We undertook extensive simulations to evaluate the existing system in a range of contrasting scenarios. We suggest several improvements relating to the treatment of trends, seasonality, re-weighting of baselines and error structure. We validate these results by running the existing and proposed new systems in parallel on real data. We find that the new system greatly reduces the number of alarms while maintaining good overall performance and in some instances increasing the sensitivity. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: negative binomial regression; outlier; outbreak; quasi-Poisson regression; robustness; statistical surveillance

1. Introduction

There has been much interest in the use of statistical surveillance systems over the last decade, prompted by concerns over bio-terrorism, the emergence of new pathogens such as SARS and swine flu, and the persistent public health problems of infectious disease outbreaks, for example the recent Escherichia coli epidemic in Germany [1]. It is important to detect these outbreaks early in order to take suitable control measures. Consequently, much statistical work has been done on this topic, inspired by a variety of statistical techniques including regression, time series models, statistical process control approaches and extensions to those fields that involve spatio-temporal studies and multivariate methods. Several reviews of recent developments have appeared in the literature [2–6].

In the UK, statistical surveillance methods have been in routine use at the Health Protection Agency (HPA) since the early 1990s and at Health Protection Scotland since the early 2000s [7, 8]. Unlike many other systems, these are designed to process very large numbers of pathogens to supplement investigator-based surveillance methods. They use a universal modelling procedure that can apply to a very diverse range of organisms with different frequencies, trends and seasonality, and that can be completely automated. The algorithm is based on a simple yet robust quasi-Poisson regression method [7]. Other public health bodies [9] have used variants of this algorithm.

One of the major challenges facing large multiple outbreak detection systems is to control a proportion of alarms that are not genuine outbreaks without impairing the detection of genuine outbreaks. A recent report on the use of the HPA outbreak detection system identified the false positive rate (FPR), namely the weekly probability of a type I error, as a limitation to its practical usefulness.

‡Supporting information may be found in the online version of this article.
In this paper, we revisit this quasi-Poisson regression-based algorithm with a view to improving its performance. Using extensive simulations, we study the sensitivity of the FPR to various modelling choices including treatment of trend, seasonality, error structure and the influence of past outbreaks. In Section 2, we describe the algorithm. Section 3 sets out the design of the simulation study, and Section 4 describes the results. In Section 5, we apply the existing and revised algorithms to the HPA data. Finally, we discuss the implications of our findings in Section 6.

2. The current algorithm and potential problems

2.1. The current algorithm

We construct the current algorithm so as to handle the broad variety of scenarios that arise in practice. We show an overall flowchart of the algorithm in Figure 1. Full details are available in [7].

![Flowchart of the overall HPA algorithm for outbreak detection of infectious diseases](image)

**Figure 1.** Flowchart of the overall HPA algorithm for outbreak detection of infectious diseases [7]. $X$ is the exceedance score, $T$ the number of reports received in the past 4 weeks and $O$ the number of reports in the current week.
description given here focuses on those aspects of the algorithm to be evaluated and possibly amended. The five key aspects of the statistical model at the heart of the algorithm are as follows.

First, we implicitly adjust for seasonal variation by basing the calculation of the expected value in the current week on the counts observed in comparable weeks in the past. Let $t$ be the current week of year $h$ and $b$ the number of years back to be considered. We take baseline counts only from weeks $t - NULNULw$ to $t - Cw$ (where $w$ is the window half-width) of years $h - b$ to $h - 1$.

Second, we include a linear trend if it is significant at the 5% level, using the log-linear model

$$\log E(y_i) = \theta + \beta t,$$

where $y_i$ is the count in week $t_i$. If the trend is not significant (or it results in an expected value outside the range of the baselines), we refit the model without it.

Third, we down-weight baseline weeks with outlying counts in order to reduce their impact on current predictions. We chose the weighting function on empirical grounds to assign very low weights to counts with large residuals. The weighting procedure is iterative (Figure 1), and the weights $w_i$ at week $t_i$ are defined by

$$w_i = \begin{cases} \gamma s_i^{-2} & \text{if } s_i > 1, \\ \gamma & \text{otherwise,} \end{cases}$$

where $\gamma$ is a constant such that $\sum_{i=1}^n w_i = n$ and the $s_i$ are scaled Anscombe residuals (see following text).

Fourth, the statistical model is a weighted quasi-Poisson model with mean $\mu_i$ and variance $\phi \mu_i$. The dispersion parameter $\phi$ is estimated by

$$\hat{\phi} = \max \left\{ \frac{1}{n - p} \sum_{i=1}^n w_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, 1 \right\},$$

where $p = 1$ or $p = 2$ depending on whether a time trend is fitted. The scaled Anscombe residuals are

$$s_i = \frac{3}{2\hat{\phi}^{1/2} \hat{\mu}_i^{1/6}} \frac{y_i^{2/3} - \hat{\mu}_i^{2/3}}{(1 - h_{ii})^{1/2}},$$

where $h_{ii}$ are the diagonal elements of the hat matrix. For Poisson data, for which $\phi = 1$, the $s_i$ are the standardised Anscombe residuals [10].

Fifth, we calculate the current expected value and the threshold value above which an observed count is declared to be unusual. The current expected count is estimated by

$$\mu_0 = \exp(\theta + \beta t_0).$$

For Poisson-distributed counts, the $2/3$ power transformation induces approximate symmetry, thus resulting in more accurate thresholds. The algorithm assumes that this applies more generally and thus that the threshold $U$

$$U = \hat{\mu}_0 \left\{ 1 + \frac{2}{3} z_{1 - \alpha} \hat{\mu}_0^{-1} (\phi \hat{\mu}_0 + \var(\hat{\mu}_0))^{1/2} \right\}^{3/2},$$

is an approximate $100(1 - \alpha)$% quantile for $y_0$, where $z_{1 - \alpha}$ is the $100(1 - \alpha)$-percentile of the normal distribution. We identify organism counts above this threshold as possible outbreaks. The output consists of a listing of organisms in order of exceedance score, defined as

$$X = \frac{y_0 - \hat{\mu}_0}{U - \hat{\mu}_0}.$$

We set the exceedance score to 0 if fewer than five reports were received in the past 4 weeks. We then flag organisms with $X \geq 1$ for more detailed investigation. The minimum outbreak size of 5 over 4 weeks reduces the likelihood that sporadic cases of infrequent organisms are flagged.

We implement this algorithm in R [11] using the package surveillance [12,13] with an overview of full content available at http://surveillance.r-forge.r-project.org.
2.2. Potential problems

In this subsection, we address possible limitations relating to the five key points listed in the preceding algorithm.

One of the limitations of the current algorithm is the small number of baseline weeks used. It is inadvisable to increase the number of years in which baseline values are taken because of long-term changes in reporting practice. The alternative is to use more of the recent data and model seasonality explicitly. Adding Fourier terms to the model to capture the broad range of seasonalities in the data is one approach. We propose to try a simple seasonal adjustment using factors. We extend the model to include a 10-level factor with a 7-week reference period (corresponding to week $t_0 \pm 3$ weeks) and further nine 5-week periods in each year:

$$\log \mu_i = \theta + \beta t_i + \delta_j(t_i),$$

where $j(t_i)$ is the seasonal factor level corresponding to week $t_i$, with $j(t_0) = 0$ and $\delta_0 = 0$, that is, the current week $t_0$ is always within the reference seasonal level.

The users of the system have reported a lack of consistency in results, which may be due to inclusion or exclusion of the trend component from week to week. We shall investigate always fitting the trend component irrespective of its statistical significance.

We are concerned that the current weighting procedure is too drastic. The baseline at week $t_i$ is down-weighted if the standardised Anscombe residual $s_i$ for that week is greater than 1. We chose this condition empirically to avoid reducing the sensitivity of the system in the presence of large outbreaks in the baselines, but this may be increasing the FPR unduly when there are no or only small outbreaks in the baselines. We will therefore investigate several other options, including reducing the down-weighting to cases where $s_i > 2$ or $s_i > 3$.

We will also evaluate a new re-weighting scheme informed by past decisions. Using this adaptive scheme, we down-weight baseline data where an alarm was flagged to reduce their effect on current predictions. The criterion we use for re-weighting, here, is the value of the exceedance score, with the following weights:

$$w_i = \begin{cases} \gamma X_i^{-1} \text{ or } \gamma X_i^{-2} & \text{if } X_i > 1, \\ \gamma & \text{otherwise}. \end{cases}$$

Finally, we will investigate the validity of the upper threshold values on the basis of the quasi-Poisson model when we generate the data using known negative binomial distributions. Although simulations alone cannot address distributional assumptions satisfactorily (as the generating model will always fit best), such studies can provide insight into the limitations of the current approach.

3. The simulation study

We first describe the simulated data and then detail the models we fit.

3.1. The simulated baseline data

We generate data using a negative binomial model (of mean $\mu$ and variance $\phi \mu$) with dispersion parameter $\phi \geq 1$, mean $\mu(t)$ and linear predictor including trend and seasonality determined by Fourier terms as follows:

$$\mu(t) = \exp \left\{ \theta + \beta t + \sum_{j=1}^{m} \left( \gamma_{1} \cos \left( \frac{2\pi j t}{52} \right) + \gamma_{2} \sin \left( \frac{2\pi j t}{52} \right) \right) \right\}.$$ 

The value $m = 0$ formally corresponds to no seasonality, $m = 1$ to annual and $m = 2$ to biannual seasonality.

To account for a wide range of data sets that arise in practice, we generate our simulations from 42 different parameter combinations taking into consideration different trends (given by values of $\beta$), seasonalities (given by values of $\gamma_{1}$ and $\gamma_{2}$), baseline frequencies of reports (given by values of $\theta$) and dispersions (given by values of $\phi$). We chose the parameter values to reflect the variety of scenarios
Table I. Parameters and criteria used to generate the 42 scenarios.

| Scenario | $\theta$ | $\beta$ | $\gamma_1$ | $\gamma_2$ | $\phi$ | $m$ | Trend |
|----------|---------|---------|------------|------------|-------|-----|-------|
| 1        | 0.1     | 0       | 0          | 1.5        | 0     | 0   |       |
| 2        | 0.1     | 0       | 0.6        | 0.6        | 1.5   | 1   | 0     |
| 3        | 0.1     | 0.0025  | 0.6        | 0.6        | 1.5   | 2   | 0     |
| 4        | 0.1     | 0.0025  | 0          | 1.5        | 0     | 1   |       |
| 5        | 0.1     | 0.0025  | 0.6        | 0.6        | 1.5   | 1   | 1     |
| 6        | 0.1     | 0.0025  | 0.6        | 0.6        | 1.5   | 2   | 1     |
| 7        | −0.2    | 0       | 0          | 0          | 2     | 0   | 0     |
| 8        | −0.2    | 0       | 0.1        | 0.3        | 2     | 1   | 0     |
| 9        | −0.2    | 0       | 0.1        | 0.3        | 2     | 2   | 0     |
| 10       | −0.2    | 0.005   | 0          | 0          | 2     | 0   | 1     |
| 11       | −0.2    | 0.005   | 0.1        | 0.3        | 2     | 1   | 1     |
| 12       | −0.2    | 0.005   | 0.1        | 0.3        | 2     | 2   | 1     |
| 13       | 1.5     | 0       | 0          | 0          | 1     | 0   | 0     |
| 14       | 1.5     | 0       | 0.2        | −0.4       | 1     | 1   | 0     |
| 15       | 1.5     | 0       | 0.2        | −0.4       | 1     | 2   | 0     |
| 16       | 1.5     | 0.003   | 0          | 0          | 1     | 0   | 1     |
| 17       | 1.5     | 0.003   | 0.2        | −0.4       | 1     | 1   | 1     |
| 18       | 1.5     | 0.003   | 0.2        | −0.4       | 1     | 2   | 1     |
| 19       | 0.5     | 0       | 0          | 0          | 5     | 0   | 0     |
| 20       | 0.5     | 0       | 0.5        | 0.5        | 5     | 1   | 0     |
| 21       | 0.5     | 0       | 0.5        | 0.5        | 5     | 2   | 0     |
| 22       | 0.5     | 0.002   | 0          | 0          | 5     | 0   | 1     |
| 23       | 0.5     | 0.002   | 0.5        | 0.5        | 5     | 1   | 1     |
| 24       | 0.5     | 0.002   | 0.5        | 0.5        | 5     | 2   | 1     |
| 25       | 2.5     | 0       | 0          | 0          | 3     | 0   | 0     |
| 26       | 2.5     | 0       | 1          | 0.1        | 3     | 1   | 0     |
| 27       | 2.5     | 0       | 1          | 0.1        | 3     | 2   | 0     |
| 28       | 2.5     | 0.001   | 0          | 0          | 3     | 0   | 1     |
| 29       | 2.5     | 0.001   | 1          | 0.1        | 3     | 1   | 1     |
| 30       | 2.5     | 0.001   | 1          | 0.1        | 3     | 2   | 1     |
| 31       | 3.75    | 0       | 0          | 0          | 1.1   | 0   | 0     |
| 32       | 3.75    | 0       | 0.1        | −0.1       | 1.1   | 1   | 0     |
| 33       | 3.75    | 0       | 0.1        | −0.1       | 1.1   | 2   | 0     |
| 34       | 3.75    | 0.001   | 0          | 0          | 1.1   | 0   | 1     |
| 35       | 3.75    | 0.001   | 0.1        | −0.1       | 1.1   | 1   | 1     |
| 36       | 3.75    | 0.001   | 0.1        | −0.1       | 1.1   | 2   | 1     |
| 37       | 5       | 0       | 0          | 0          | 1.2   | 0   | 0     |
| 38       | 5       | 0       | 0.05       | 0.01       | 1.2   | 1   | 0     |
| 39       | 5       | 0       | 0.05       | 0.01       | 1.2   | 2   | 0     |
| 40       | 5       | 0.0001  | 0          | 0          | 1.2   | 0   | 1     |
| 41       | 5       | 0.0001  | 0.05       | 0.01       | 1.2   | 1   | 1     |
| 42       | 5       | 0.0001  | 0.05       | 0.01       | 1.2   | 2   | 1     |

arising in the HPA surveillance data [14]. We give the parameters for the 42 scenarios in Table I. Figure 2 shows four of the resulting data series.

The simulations of the baseline data (that is, in the absence of outbreaks) use 100 replicates from each scenario, each of size $n = 624$ weeks. We used the last 49 weeks as ‘current’ weeks; thus, the simulation results are based on $100 \times 49 = 4900$ replicates for each of the 42 data scenarios and each model investigated (we describe the models in Section 3.4).

We use weeks 1–312 as training weeks for the adaptive re-weighting schemes and weeks 313–575 as baseline weeks. We use weeks 576–624 as the ‘current’ weeks.
3.2. The simulated outbreaks

We simulated an outbreak starting in week $t_i$ as follows. First, we selected a given constant $k$ and randomly generated the outbreak size as Poisson with mean equal to $k$ times the standard deviation of the baseline count at $t_i$. We then randomly distributed these outbreak cases in time according to a lognormal distribution with mean 0 and standard deviation 0.5. Values of $k$ in the range 2–10 typically resulted in outbreak durations of 2–8 weeks. We added the outbreak cases in week $z$ after the start of the outbreak to the baseline counts in week $t_i + z$. We recorded the start and end times of the outbreak. We added outbreaks to baseline and current weeks as follows.

Outbreaks in baseline weeks. For each data series, we included four outbreaks with start times chosen randomly among the baseline weeks (weeks 313–575). We took the values of $k$ to be 2, 3, 5 or 10. We generated different outbreaks for each of the 4900 runs.

Outbreaks in current weeks. For each data series, we included one outbreak with start time chosen randomly among the last 49 weeks (weeks 576–624). We took the values of $k$ to be in the range 1–10. We generated different outbreaks for each of the 4900 runs.

3.3. Evaluation measures

We used different measures to evaluate the performance of the detection system in the absence and presence of current outbreaks in order to capture quantities that are meaningful in the present context.

When there are no outbreaks in the baselines, we calculated the FPR for each of the 42 scenarios as the proportion of the current 49 weeks and 100 replicates, in which the observed value exceeded the threshold in the absence of any current outbreak in those 49 weeks.

Figure 2. Plots of simulated data series (a) scenario 8, (b) scenario 10, (c) scenario 12 and (d) scenario 17 (see Table I).
To evaluate the impact of outbreaks in the baselines, we calculated the FPR as earlier but only included in the calculation those current weeks for which at least one baseline week coincided with a baseline outbreak.

We calculated the probability that an outbreak is detected (POD), or power, for each of the 42 scenarios as follows. We ran the algorithm iteratively for the 49 current weeks, and if the threshold was exceeded at least once between the start and end times of the outbreak, we deemed the outbreak to have been detected. We then calculated the POD as the proportion of outbreaks detected in 100 runs. We did not restrict the calculation when evaluating the impact of outbreaks in the baselines.

Note that the FPR is a rate ‘per week’, whereas the POD is a rate ‘per outbreak’.

We use these evaluation measures because they relate to the performance of the system on individual time series. In Section 6, we return to the issue of which criteria to use when the system is in operation on real data, including the use of measures such as the false discovery rate (FDR).

3.4. The models

We evaluated a sequence of algorithms starting with a slightly simplified version of the HPA algorithm. All algorithms used 5 years baseline values ($b = 5$ years) and half-windows of 3 weeks ($w = 3$ weeks); thus, for the HPA method, we used 35 baseline values. We took the current counts from the final 49 weeks of the simulated series. We used thresholds based on the 0.995 and 0.975 quantiles.

Figure 3. False positive rates obtained using Model 0 with thresholds based on 0.995 (a) and 0.975 (b) quantiles. The dashed lines represent the nominal values. Numbers refer to scenarios (see Table I).
Model 0 (The HPA model)
We used a slight modification of the algorithm depicted in Figure 1. Thus, we removed the condition on minimal outbreak size (namely that outbreaks are only flagged if the total organism count in the most recent 4 weeks is at least 5). We also removed the condition specifying that no trend is fitted if the expected current value with a trend lies outside the range of the 35 baseline counts.

Model 1
This model is the HPA model (Model 0) without any re-weighting. It provides a baseline from which to evaluate different ways of modelling trend and seasonality. We will consider alternative weightings in later models.

Model 2
This model is the same as Model 1, except that we always fit a trend regardless of its statistical significance. We chose this to avoid the inconsistencies resulting from switching between including and not including marginally significant trends in adjacent weeks.

Model 3
Here, we extend Model 2 to include seasonal factors. Thus, we make use of all data in the past 5 years up to 26 weeks before the current week. The reason we exclude the last 26 weeks from the baseline is to avoid adaptation of the model to emerging outbreaks which would reduce its sensitivity. We used 10-level factors including a 7-week reference period and further nine 5-week periods in each year. We chose this model to improve the estimation of trend and dispersion.

Model 4
This model is the same as Model 3 except that the threshold for detecting outbreaks and for calculating the exceedance score is based on the negative binomial distribution rather than the normal approximation.

Figure 4. False positive rates obtained using Models 1 (a), 2 (b), 3 (c) and 4 (d) with threshold based on the 0.995 quantile. The dashed lines represent the nominal values. Numbers refer to scenarios (see Table I).
involving scaled Anscombe residuals. It is of interest to quantify the loss of accuracy resulting from this normal approximation.

We fitted the preceding sequence of models to the simulated data without outbreaks to evaluate the FPR. Models 1–4 do not include any re-weighting of high values in baselines. Such re-weighting is required to avoid reducing the sensitivity of the algorithm when there are outbreaks in the baseline weeks. In further simulations, we elaborated Model 4 to incorporate several different weighting schemes and evaluated them first on outbreak-free data, as we now describe.

Model 5
This model is the same as Model 4, with a re-weighting scheme similar to that used in Model 0, on the basis of the scaled Anscombe residuals \( s_i \) as described in Section 2.2. Thus, there are three versions of this model according to whether baseline counts are re-weighted when \( s_i > 1 \), \( s_i > 2 \) or \( s_i > 3 \).

Model 6
This model is also the same as Model 4 but with the adaptive re-weighting scheme described in Section 2.2. We evaluated four versions of this model, using exceedance scores \( X \) in weeks 313–575 based on the 0.995 and 0.975 quantiles and re-weighting factors \( X^{-1} \) and \( X^{-2} \). We obtained the exceedance scores \( X \) for the baseline weeks by running Model 4 with ‘current’ weeks 313–575 (and thus using baseline weeks 1–312 uncontaminated by outbreaks).

We also investigated the performance of Models 5 and 6 in the presence of outbreaks of varying sizes in the baselines as described in the previous section.

Finally, we investigated the power of Models 5 and 6 to detect current outbreaks of \( k = 1–10 \) standard deviations, both with \( (k = 3 \) and \( 10) \) and without outbreaks in the baselines.

Figure 5. False positive rates obtained using Model 5 with re-weighting procedure based on scaled Anscombe residuals, \( s_i > \infty \) (a), \( s_i > 3 \) (b), \( s_i > 2 \) (c) and \( s_i > 1 \) (d) with threshold based on the 0.995 quantile. The dashed lines represent the nominal values.
4. Results of the simulations

4.1. False positive rates

Figure 3 shows the FPRs obtained using Model 0 with thresholds based on 0.995 and 0.975 quantiles. We show all FPRs as central estimates and 95% confidence intervals. The nominal FPRs are 0.005 and 0.025. The actual FPRs are much larger, confirming concerns that the HPA algorithm produces too many false alarms. Performance is good only when the data are Poisson. It is particularly bad for scenarios involving a steep trend rising from a low baseline with over-dispersed data (e.g. scenarios 10–12). The effect of seasonality appears to be relatively small.

Figure 4 shows the FPRs obtained using Models 1–4 with a threshold based on the 0.995 quantile (results relating to the 0.975 quantile are in the Supplementary Material). Both figures show monotone improvements as we go from Model 1 through to Model 4. The improvement is particularly marked when going from Model 3 to Model 4 with nominal FPR 0.005 (Figure 4). Introducing factors in Model 3 further decreases the FPR for most scenarios. However, they introduce some additional bias for scenarios with very high amplitude of seasonal variation (e.g. scenarios 27 and 30).

Figures 5 and 6 show the impact on Model 4 of incorporating weights in the baselines in the absence of any outbreaks, using thresholds based on the 0.995 quantile. Figure 5 shows the impact of weighting procedures based on scaled Anscombe residuals $s_i = \infty$, $s_i > 3, 2, 1$, the first one corresponding to no re-weighting and the last to the weighting used in the HPA algorithm. The performance clearly deteriorates as the weighting increases, as expected. Figure 6 shows the impact of weighting procedures based on the adaptive schemes with weighting factors $X^{-1}$ and $X^{-2}$, with $X = 1$ corresponding to 0.995 and 0.975 quantiles. Generally, the impact is less than with the Anscombe weights. There is a slight deterioration in performance with weighting factor $X^{-2}$ and 0.975 quantile.

![Figure 6](image_url)
The effect of re-weighting on reducing the impact of outbreaks in the baselines is shown in Figures 7 and 8 for outbreaks of 3 and 10 standard deviations, respectively. We show the Anscombe-based schemes with $s_i > 2$ and $s_i > 3$ and the adaptive schemes based on 0.975 quantiles (other scenarios including other re-weighting schemes and outbreaks of two and five standard deviations are in the Supplementary Material). For outbreaks of three standard deviations, the stronger re-weighting schemes ($s_i > 2$, $X_{NUI}^{-1}$) still slightly increase the FPR. For outbreaks of 10 standard deviations, the weaker schemes ($s_i > 3$, $X^{-2}$) reduce the FPR to well below its nominal value.

4.2. Probability an outbreak is detected

Figure 9 shows the power curves for Model 5 with re-weighting schemes $s_i > 2$ and $s_i > 3$ in the presence of outbreaks of three standard deviations in the baselines. We also show the pointwise median of the 42 power curves. As expected, the PODs increased with $k$ and were higher for re-weighting on the basis of $s_i > 2$ than $s_i > 3$, the largest difference being at $k = 5$. The spread of the PODs across the 42 scenarios was less with the $s_i > 2$ weighting scheme. We obtained similar results with no outbreaks in the baselines and outbreaks of 10 standard deviations in the baselines (Supplementary Material). We summarise all results in Table II, which shows the median PODs across the 42 scenarios.

Some individual scenarios produce unusual power curves. For example, scenarios 27 and 30 have consistently lower PODs, mirroring the low FPR also observed for these scenarios, which correspond to highly seasonal baselines. On the other hand, series 7–9 resulted in high PODs at small current outbreak sizes. These scenarios have very low baseline values.

Figure 10 shows the power curves for Model 6 with the adaptive re-weighting schemes with thresholds based on the 0.975 quantiles and re-weighting factors $X^{-1}$ and $X^{-2}$ in the presence of outbreaks.
Figure 8. False positive rates obtained using Model 5 for outbreaks of 10 standard deviations with re-weighting procedure based on scaled Anscombe residuals $s_i > 3$ (a) and $s_i > 2$ (b) with threshold based on the 0.995 quantile and using Model 6 with weighting procedures based on the adaptive schemes with weighting factors $X^{-1}$ (c) and $X^{-2}$ (d) and threshold based on the 0.975 quantiles. The dashed lines represent the nominal values.

Figure 9. Proportions of detected outbreaks of $k$ standard deviations (along with the median curves in bold) corresponding to the 42 scenarios. Outbreaks of three standard deviations are included in baselines. The re-weighting procedure is based on scaled Anscombe residuals $s_i > 3$ (a) and $s_i > 2$ (b) with threshold based on the 0.995 quantile.

of three standard deviations in the baselines. The PODs are similar to those in Figure 9. Specifically, the adaptive scheme based on $X^{-1}$ produces similar powers to that based on $s_i > 3$, and the scheme based on $X^{-2}$ produces only slightly lower powers than that based on $s_i > 2$. This is confirmed in Table III.
Table II. Median proportions of detected outbreaks of $k$ standard deviations across the 42 scenarios with the Anscombe-based re-weighting procedures with $s_i > 2$ and $s_i > 3$.

| Current outbreak size | Anscombe Baseline re-weighting outbreak schemes |
|-----------------------|-----------------------------------------------|
|                       | Baseline outbreak size | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
| $s_i > 2$             | 0 | 0.050 | 0.120 | 0.220 | 0.340 | 0.465 | 0.610 | 0.750 | 0.840 | 0.905 | 0.950 |
|                       | 3 | 0.040 | 0.090 | 0.190 | 0.330 | 0.460 | 0.590 | 0.715 | 0.830 | 0.910 | 0.960 |
| $s_i > 3$             | 0 | 0.040 | 0.080 | 0.145 | 0.290 | 0.375 | 0.520 | 0.650 | 0.780 | 0.865 | 0.905 |
|                       | 3 | 0.030 | 0.070 | 0.130 | 0.245 | 0.320 | 0.490 | 0.640 | 0.720 | 0.845 | 0.925 |
|                       | 10 | 0.020 | 0.040 | 0.080 | 0.140 | 0.245 | 0.315 | 0.435 | 0.560 | 0.700 | 0.820 |

Figure 10. Proportions of detected outbreaks of $k$ standard deviations (along with the median curves in bold) corresponding to the 42 scenarios. Outbreaks of three standard deviations are included in baselines. The re-weighting procedure is based on the adaptive schemes with weighting factors $X^{-1}$ (a) and $X^{-2}$ (b) and threshold based on the 0.975 quantiles.

Table III. Median proportions of detected outbreaks of $k$ standard deviations across the 42 scenarios with the adaptive schemes with weighting factors $X^{-1}$ and $X^{-2}$.

| Current outbreak size | Anscombe re-weighting scheme | Baseline outbreak size | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|-----------------------|--------------------------------|------------------------|----|----|----|----|----|----|----|----|----|----|
|                       | $X^{-1}$                       |                        | 0  | 0.040 | 0.080 | 0.150 | 0.250 | 0.345 | 0.500 | 0.695 | 0.790 | 0.885 | 0.940 |
|                       |                               | 3 | 0.030 | 0.070 | 0.125 | 0.230 | 0.380 | 0.485 | 0.620 | 0.790 | 0.825 | 0.935 |
|                       |                               | 10 | 0.010 | 0.030 | 0.080 | 0.150 | 0.200 | 0.325 | 0.440 | 0.590 | 0.755 | 0.805 |
|                       | $X^{-2}$                       |                        | 0  | 0.035 | 0.090 | 0.180 | 0.270 | 0.380 | 0.580 | 0.695 | 0.805 | 0.890 | 0.945 |
|                       |                               | 3 | 0.040 | 0.080 | 0.170 | 0.270 | 0.345 | 0.515 | 0.700 | 0.780 | 0.865 | 0.920 |
|                       |                               | 10 | 0.025 | 0.050 | 0.110 | 0.235 | 0.335 | 0.435 | 0.590 | 0.730 | 0.875 | 0.890 |

5. Evaluations on data

The simulations in the previous section show that the adaptive down-weighting schemes do not perform substantially better than those based on scaled Anscombe residuals. As they are more difficult to implement in practice, we shall not consider them further. The simulations also show that the current HPA system inflates the FPR and that Model 5 with re-weighting based on $s_i > 2$ or 3 is a substantial improvement. Hereafter, we shall use a compromise between these two re-weighting schemes and use $s_i > 2.58$, so that the 0.995 quantile is used both for the re-weighting and for the aberration threshold.
The simulations provide little guidance on whether one should use a quasi-Poisson or a negative binomial model with real data. In this section, we compare the performance of the original HPA model (Model 0) and Models 3 (quasi-Poisson) and 4 (negative binomial) with re-weighting using $s_i > 2.58$ on real data.

For current data, we used weekly reports of organism counts for 2011 from the HPA Lab-Base surveillance data base with baseline counts drawn from the previous 5 years, with sequences of leading zeroes recoded as missing.

In the first analysis, we ran the three algorithms in parallel on these 52 weeks’ data. We show the total numbers of organisms flagged as aberrant with the three algorithms in Table IV. Because the current HPA system does not report aberrances unless the total number of organisms reported in the past 4 weeks is 5 or more, Table IV also shows the numbers of aberrances with this filter. Under the current system, some 18 organisms are classified as aberrant on average each week (not counting those with zero baselines). The new systems reduce this to under 10. We calculated detection rates using a denominator comprising the number of organism-weeks for which an expected value could be calculated under each system (namely at least one nonzero baseline count and at least 2 years’ baseline data) and which

| Table IV. Numbers and rates per hundred organism-weeks (among those with nonzero baselines) of aberrant organisms in 2011 obtained by the three systems, both including and excluding those with $T < 5$. |
|---------------------------------------------------------------|
| **HPA** | **Quasi-Poisson** | **Negative binomial** |
| Including $T < 5$ Numbers aberrant | 1244 | 743 | 591 |
| Rate | 2.507 | 0.867 | 0.689 |
| Excluding $T < 5$ Numbers aberrant | 924 | 497 | 416 |
| Rate | 1.862 | 0.580 | 0.485 |

HPA, Health Protection Agency.

| Table V. Aberrances for week 22 under Models 0 (HPA model), 3 (quasi-Poisson) and 4 (negative binomial). |
|---------------------------------------------------------------|
| **Organism** | $y_0$ | $T$ | **HPA** | **Quasi-Poisson** | **Negative binomial** |
|---------------|-------|-----|--------|------------------|------------------|
| *Brucella abortus* | 2 | 2 | 1.57 | 1.04 |
| *Chlamydia* other named | 34 | 80 | 1.22 |
| *Corynebacterium striatum* | 4 | 7 | 1.08 |
| *Escherichia coli* | 2 | 2 | 1.29 |
| *Escherichia coli* O 157 PT 8 VT2 | 40 | 175 | 2.34 | 1.30 | 1.20 |
| *Fusarium oxysporum* | 3 | 4 | 1.80 | 1.01 |
| *Haemophilus* other named | | | 3.30 | 3.30 |
| *Klebsiella* sp. | 147 | 204 | 1.04 |
| *Lactococcus* sp. | 2 | 2 | 1.08 |
| *Legionella pneumophila* ungrouped | 2 | 3 | 1.05 |
| *Microsporum* sp. | 1.48 | 2.38 |
| *Salmonella enteritidis* PT 2 | 3 | 6 | 1.41 | 1.12 |
| *Salmonella* give | 3 | 3 | 1.06 | 1.25 | 1.39 |
| *Salmonella minnesota* | 3 | 4 | 1.52 | 1.54 | 1.36 |
| *Salmonella montevideo* | 9 | 28 | 2.91 | 2.26 | 2.63 |
| *Salmonella newport* | 10 | 21 | 1.04 | 1.01 |
| *Salmonella typhimurium* DT 2 | 3 | 4 | 1.89 | 1.12 |
| *Salmonella typhimurium* DT 40 | 2 | 6 | 1.28 | 1.60 | 1.64 |
| *Shigella boydii* 14 | 1.59 | 1.59 |
| *Shigella sonnei* PT 6 | 9 | 36 | 1.79 | 1.73 | 1.72 |
| *Shigella sonnei* RDNC | 3 | 8 | 2.09 | 1.76 | 2.28 |
| *Stenotrophomonas* sp. | 1.59 | 1.59 |
| *Trichophyton* other named | 11 | 51 | 1.20 |

For each organism aberrant on at least one model, the observed count $y_0$, the total count in the most recent 4 weeks $T$ and the exceedance score $X_2$, calculated under Equation (2), are shown.

HPA, Health Protection Agency.
varies between methods owing to the different numbers of baseline weeks used. Note that this choice of
denominator is by no means unique. These rates are shown in Table IV.

Table IV shows that the new systems reduce the detection rate by about a factor of 3 for the quasi-
Poisson model and a little more for the negative binomial. The simulations suggest that the difference is
due to a reduction in the FPR.

To examine these results in greater detail, we focus on aberrances for a single week, week 22. We also
investigated week 52 with similar results, not reported here. In week 22, 391 organisms were reported.

Table V shows the aberrances for week 22 under the three models. For each organism aberrant on at
least one model, we show the observed count $y_0$ and the total count in the most recent 4 weeks, $T$. We
also show the exceedance scores, calculated using the new formula

$$X_2 = \frac{y_0^{2/3} - \hat{\mu}_0^{2/3}}{U^{2/3} - \hat{\mu}_0^{2/3}}.$$  

(2)

We used this formula rather than that in Equation (1) because it is on the same scale as the quantiles.

There were 19 aberrances under the HPA model, 16 under the quasi-Poisson model and 11 under the
negative binomial. The HPA model flagged seven organisms that were not flagged by the other two. The
HPA and quasi-Poisson flagged five organisms that were not flagged by the negative binomial. Interestingly, both new methods flagged four organisms that were not flagged by the version of the HPA model
which we implemented. These correspond to organisms with zeroes in all baselines. Because the new
methods use past counts outside the baseline values, it is possible to calculate prediction intervals for
these organisms. Under the existing system, they are just given a default exceedance score of 1 (we did
not implement this here), but no quantification of the aberrance is possible.

Figure 11 shows the plots for four organisms with discrepant results. The HPA model alone flagged the
top two. For both, the expected value under the HPA system was low (2.07 for *Chlamydia* other named
and 29.25 for *Klebsiella* sp.). *Haemophilus* other named has zero baselines and so the HPA system did

---

Figure 11. Plots for *Chlamydia* other named (a), *Klebsiella* sp. (b), *Haemophilus* other named (c) and *Salmonella* enteritidis PT 2 (d).
not flag it in spite of the highly aberrant current count. The HPA and quasi-Poisson models flagged *Salmonella enteritidis* PT 2, which has low counts, but the negative binomial model did not flag it, most likely because of this model avoiding inaccurate normal approximations. These examples suggest that the new systems are producing more reasonable results.

6. Discussion

We have undertaken a thorough evaluation of the HPA’s outbreak detection system on the basis of simulated and real data. The main conclusion from this evaluation is that the FPR is too high, owing to a combination of factors, notably excessive down-weighting of high baselines and reliance on too few baseline weeks. We have studied several alternatives that reduce the FPR without seriously degrading the POD. One set of alternatives used an adaptive re-weighting scheme. We found this to give broadly equivalent results to the re-weighting method based on scaled Anscombe residuals. Our recommendation is to use the latter as in the original HPA method but with much higher threshold for re-weighting, namely, \( s_i > 2.58 \) rather than \( s_i > 1 \).

A novel feature of these new models is that they make use of much more baseline data. This means that the trend and variance are better estimated. In addition, we recommend that the trend should always be fitted even when non-significant (or extreme). This will decrease the discrepancies in the results when moving from one week to another.

One might think that the dispersion could vary with the mean. Thorough empirical analyses of HPA data over 20 years [14] suggest that this is not a serious problem. These investigations also suggest that the negative binomial model is a reasonable one, although not ideal in all circumstances. Thus, there is a good case for replacing the quasi-Poisson model with the negative binomial. This has also been recommended in [15]. However, the overall results we obtained on the 2011 data showed only moderate differences between the numbers of organisms flagged by these models.

One of the unusual features of the HPA system is that it is run every week on a database of more than 3300 distinct organisms, which is likely to produce a large number of aberrances. We recommend retaining the exceedance score approach based on the 0.995 quantile using \( X_2 \) from Equation (2). This involves ranking aberrant organisms in order of exceedance. This system combines the week-to-week comparability of threshold-based results with the operational flexibility of limiting investigations to the most serious aberrances. An alternative would be to use a fixed maximum FDR, say. This would work as follows. In a given week, the \( m \) ordered exceedance scores are replaced by their associated \( p \)-values, which are therefore in increasing order, and the list of potential outbreaks is truncated at the last point at which the \( i \)th \( p \)-value does not exceed \( (i/m) \) [16]. There are two problems with this approach. First, we may need to adapt the method to condition on there being at least one report. More problematically, however, the outbreak threshold will vary from week to week, which can produce apparently inconsistent results for the same organism in adjacent weeks. Because epidemiologists use information from recent weeks to assess current potential outbreaks, such week-to-week variation in the outbreak threshold as applied to an individual organism may prove unhelpful. This feature of the FDR has proved problematic in other applications [17].

Acknowledgements

This research was supported by a project grant from the Medical Research Council and by a Royal Society Wolfson Research Merit Award.

References

1. Buchholz U, Bernard H, Werber D, Böhmer MM, Remschmidt C, Wilking H, Deleré Y, an der Heiden M, Adlhoch C, Dreesman J, Ehlers J, Ethelberg S, Faber M, Frank C, Fricke G, Greiner M, Höhle M, Ivarsson S, Jark U, Kirchner M, Koch J, Krause G, Luber P, Rosner B, Stark K, Kühne M. German outbreak of *Escherichia coli* O104:H4 associated with sprouts. *The New England Journal of Medicine* 2011; 365:1763–1770.
2. Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A* 2003; 166:5–21.
3. Farrington CP, Andrews N. Outbreak detection: application to infectious disease surveillance. In *Monitoring the Health of Populations: Statistical Principles & Methods for Public Health Surveillance*, Brookmeyer R, Stroup DF (eds). Oxford University Press: Oxford, 2004; 203–231.
4. Shmueli G, Burkom HS, Campbell M, Hogan WR, Moore A. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics* 2005; **38**:99–113.

5. Buckeridge DL, Burkom HS. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics* 2010; **52**:39–51.

6. Unkel S, Farrington CP, Garthwaite PH, Robert CS, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society Series A* 2012; **175**:49–82.

7. Farrington CP, Andrews NJ, Beale AJ, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society Series A* 1996; **159**:547–563.

8. McCabe GJ, Greenhalgh D, Gettingby G, Holmes E, Cowden J. Prediction of infectious diseases: an exception reporting system. *Journal of Medical Informatics and Technologies* 2003; **5**:67–74.

9. Hulth A, Andrews N, Ethelberg S, Dreesman J, Faensen D, van Pelt W, Schnitzler J. Practical usage of computer-supported outbreak detection in five European countries. *Eurosurveillance* 2010; **15**(36):8–13.

10. Davison AC, Snell EJ. Residuals and diagnostics. In *Statistical Theory and Modeling*, Hinkley DV, Reid N, Snell EJ (eds). Chapman and Hall: London, 1991; 83–106.

11. R Development Core Team. R: a language for and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. http://www.R-project.org

12. Höhle M. Surveillance: an R package for the monitoring of infectious diseases. *Computational Statistics* 2007; **22**:571–582.

13. Höhle M, Mazick A. Aberration detection in R illustrated by Danish mortality monitoring. In *Biosurveillance: Methods and Case Studies*, Kass-Hout T, Zhang X (eds). CRC Press: London, 2010; 215–237.

14. Enki GE, Noufaily A, Garthwaite PH, Andrews NJ, Charlett A, Farrington CP. Twenty years of automated biosurveillance: data from the England and Wales outbreak detection system. *Technical Report 12/05*, Open University Statistics Group, 2012. Available from statistics.open.ac.uk.

15. Wagner A. Extending the Scottish exception reporting systems spatially and temporally. *PhD Thesis*, University of Strathclyde, 2010.

16. Jones HE, Ohlssen DI, Spiegelhalter DJ. Use of the false discovery rate when comparing multiple health care providers. *Journal of Clinical Epidemiology* 2008; **61**:232–240.

17. Spiegelhalter D, Sherlow-Johnson C, Bardsley M, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society Series A* 2012; **175**:1–47.