**Supplemental Appendix 3: R Shiny Quantitative Calculations**

The R Shiny application takes an excel, csv, or sas7bdat dataset file aggregated into Date (month-day-year form) and Proportion (in % of tests above the blood lead level threshold) columns or with the relevant columns for aggregating. The R Shiny app outputs alert levels for each month in both table and visual form, as detailed in the following flowchart:



This process involves splitting the data into windows and performing the analyses on each window to indicate the alert level for the final month of that window. By default, windows are 61 months wide (unless 5 years of data isn’t available) and the 61st month (or final month) is the only one to get an alert level; this alert is collected across all windows to create the output table and visuals. The processes for splitting and analyzing are further defined mathematically below:

## Data Splitting

Given the primary dataset with, for example, 108 months of data aggregated at the proportion of tests ≥ 5µg/dL within a given month:

|  |  |  |
| --- | --- | --- |
| Index [1 – 108] | Date [in Month-Year format] | Proportion of Tests ≥ 5*µg/dL* |
| 1 | 01-2011 | 0.11 |
| 2 | 02-2011 | 0.13 |

98 **secondary datasets {,** for i = 1, 2, 3, …, 98**}** are created as subsets of the **primary** datasetfor analysis, as different slices of the index, starting with a widening window for the first 61 months, then a rolling window for the subsequent 48 slices of 61 months each. This is designed so data with less than the 5 years of data baseline can still be analyzed. In the R Shiny application, the initial window (here 11) and the window size (here 61) can be changed by the user.

Index items {Number of rows in dataset }

1 – 11 {N = 11}

1 – 12 {N = 12}

1 – 13 {N = 13}

…

1 – 60 {N = 60} [50 datasets so far]

1 – 61 {N = 61}

2 – 62 {N = 61}

3 – 63 {N = 61}

…

48 – 108 {N = 61} [48 more datasets]

This allows every month of the **primary dataset** (except for the first 10) to be the final month of its own **secondary dataset**, a requirement for the CUSUM/Shewhart being that a single month is out of quality control compared to only the months before it. Unlike other time series models, such as seasonal adjustment via LOESS, we do not want the value Jan 2015 to affect the residuals for Jan 2013.

## Modeling the Residual

To account for potential correlation over time and the two recognized patterns (long term downward trend and seasonality peaking in late summer or early fall), we fit an autoregressive-1 model for each **secondary dataset** that includes a linear time trend, with a join point that allows the pattern to change. To account for the seasonal pattern, we also include linear time splines with seasonal knots.

The Model is of the linear form:

Where the error term is assumed to be normally distributed:

|  |  |
| --- | --- |
| Variable | Definition |
|  | Proportion of Tests ≥ 5*µg/dL* in month *t*. |
|  | The intercept value. |
|  | Time of month *t* as defined by what row number it is in its current dataset. This will at most equal 61, because that is the largest any **secondary** dataset will be.  |
|  | The 36th month join point, defined as the number of months since *t = 36*. [0, 0, 0, 0... (36 zeroes total)… 1, 2, 3…] For **secondary** datasets with less than 40 months, this column is not used.  |
|  | Number of months since January (this year), e.g. [1, 2, 3, … 12] repeating.  |
|  | Number of months since March (this year), e.g. [0, 0, 0, 1, 2, 3… 9] repeating. |
|  | Number of months since June. (this year), e.g. [0, 0, 0, 0, 0, 0, 1, 2, 3 … 6] repeating. |
|  | Number of months since September (this year), e.g. [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 3] repeating. |
|  | Proportion of Tests ≥ 5*µg/dL* in month (*t* – 1). The autoregressive component. |

The coefficients for this model are determined using Maximum Likelihood as the estimation method and the Broyden–Fletcher–Goldfarb–Shanno Algorithm (BFGS) as the optimization method.

The splines are constrained so that the spline value at months 1 and 13 will coincide (a 12-month period).

Thus, using:

The constrained Autoregressive Model becomes:

Where:

We fit each of the 98 **secondary dataset** models using (N – 1) data points and calculate the model-predicted value for month N (the most recent BLL in the **secondary dataset,** i.e. the 61st month for all the windows with 5+ years of data).

## Residuals and Standard Errors

The residuals,, are used to calculate standard errors for each point.

Where is the autoregressive-adjusted t row of the design matrix **X**:

(Recall, \* column excluded if N < 40.)

That is, the design matrix does not include the autoregressive term (whose coefficient is) so the previous row of the design matrix is scaled by and added. This matrix is then multiplied by the variance-covariance matrix of the (constrained) non-autoregressive parameters ():

And then multiplied by its transpose which returns a matrix to be added to

Where is the Mean Square Error (MSE) with degrees of freedom ():

We fix , because of our assumption that . Without this assumption, *r* would be defined through a process on (see the Methods section for more details on this decision). The differences made by this assumption have been negligible using SAS.

Thus, .

These standard errors are used to calculate variance value for every data point in every **secondary dataset.**

The median *v* value () within each **secondary dataset {,** for i = 1, 2, 3, …, 98**}** is used to determine the standard deviation used in the CUSUM/Shewhart calculation for that secondary dataset (i.e. that window).

## CUSUM Calculation

Each month in each **secondary dataset** gets a CUSUM value:

Where the reference interval (k) was determined exogenously to be 1 (see the SAS defaults in the Methods section) and the z-score ():

Where the target mean was exogenously chosen to be 0 (see the SAS defaults in the Methods section).

The decision interval (h) was determined exogenously to be 3 (see the SAS defaults in the Methods section), but instead of a binary exceeded/ not-exceeded framework, we labeled each as one of the following classifications based on its value:

 3+ st.dev

 2+ st.dev

 1+ st.dev

 Else < 1 st.dev

Plotting all 108 months, where the first 10 months received the Else value by default, and the latter 98 months were colored based on the values for each **secondary dataset** where (i.e. that month was the final month of its own **secondary dataset**). See Figure 4, duplicated here:



## Shewhart Calculation

The Shewhart Irchart method uses the moving range of the residuals to determine an outlier; where the moving range is defined as:

For t in 2 to N:

Both the direct average and the absolute expected value of are calculated to determine the lower confidence limit (not relevant for our analysis) and the upper confidence limit for the residual. Note that is excluded in the absolute expected value, *MR*.

Where the limiting multiplier value 2.66 is obtained by dividing 3 by the sample size-specific d2 anti-biasing constant for n=2 (note: this is the SAS default). Montgomery, Douglas (2005). Introduction to Statistical Quality Control. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN 978-0-471-65631-9. OCLC 56729567. Archived from the original on 2008-06-20.

The Shewhart UCL value is used to make the warning line (red line in Figure 4) and the value is treated as binary. If the residual exceeds the UCL it is given a 3+ std. dev. value (red) regardless of its cusum value; otherwise, it is given its cusum value.

Appendix 3 References

Montgomery, Douglas (2005). Introduction to Statistical Quality Control. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN 978-0-471-65631-9. OCLC 56729567. Archived from the original on 2008-06-20.