

Better Practice through Advanced Analytics

2:30PM EST | November 18th, 2021 | 2021 SyS Symposium

Moderator: Sara Chronister, MPH – Maricopa County Department of Public Health (AZ)

Presenters:

Hannah P. Cowley, BA – Johns Hopkins University Applied Physics Laboratory

Michael Sheppard, MS – CDC/NSSP

Howard Burkom, PhD – Johns Hopkins University Applied Physics Laboratory

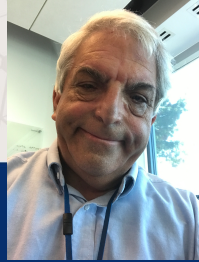


Council of State and Territorial Epidemiologists



Hannah P. Cowley, B.A.
Johns Hopkins University Applied
Physics Laboratory





Michael Sheppard, MS – CDC &
Howard Burkom, PhD - Johns Hopkins
University Applied Physics Laboratory



Analytic Discriminator to Avoid Algorithm Alerts Resulting from Data Quality Problems

Howard Burkom, PhD

Mathematician, Johns Hopkins University Applied Physics Laboratory

Michael Sheppard, MS

Mathematical Statistician, Data Science and Innovation Team
Surveillance and Data Branch, Division of Health Informatics and Surveillance
Center for Surveillance, Epidemiology, and Laboratory Services
CDC

Roseric Azondekon, PhD

Health Scientist, Data Science and Innovation Team
Surveillance and Data Branch, Division of Health Informatics and Surveillance
Center for Surveillance, Epidemiology, and Laboratory Services
CDC



November 18, 2021

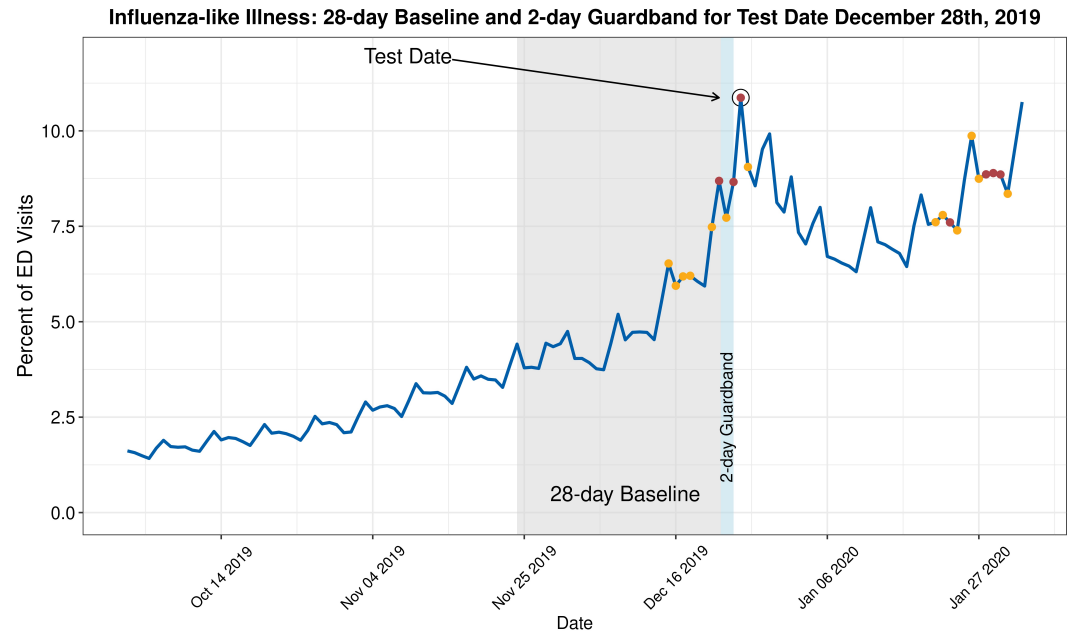
Issues of General Concern in Surveillance

- Avoiding signals caused by data problems, not events of concern
- Generating truth data
 - Expert annotation is expensive & time-consuming
- Generalizing, extending labelled data
- Calibration using labelled data

Basic Problem: Are Data Anomalies Results of a Detected Event or Data Quality Issues?

Recognition of the abnormal requires understanding of expected data behavior

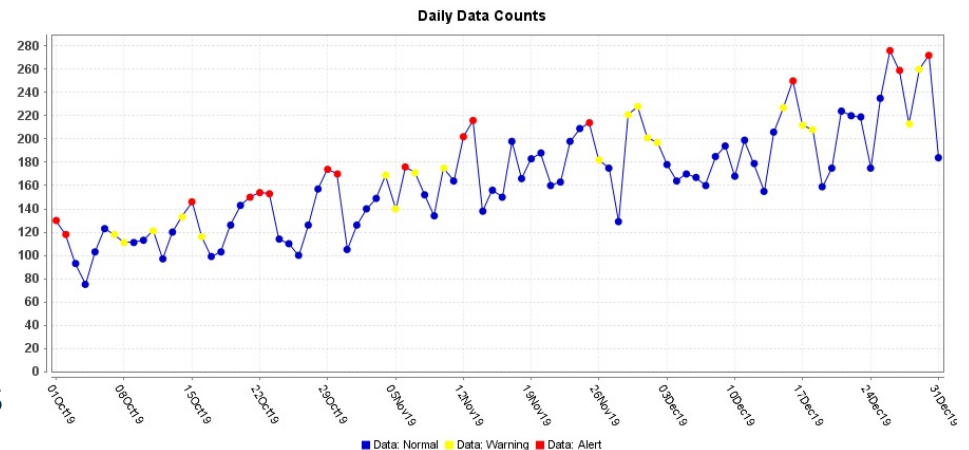
- Expectations need to be updated in many applications
- In ESSENCE systems, we use a sliding 28-day baseline with a 2-day guardband to separate the baseline from the test date



ESSENCE Datasets Used

From counties with clearest data quality problems:

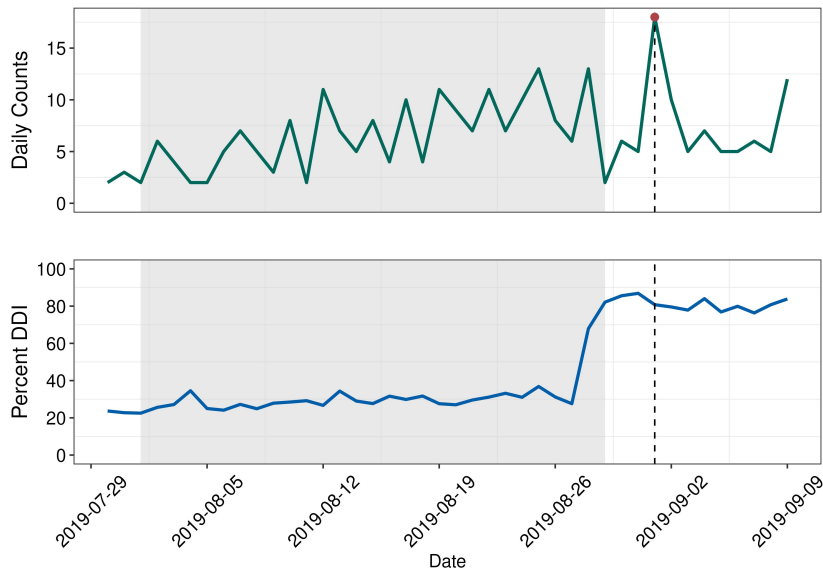
- Time series of daily counts of influenza-like-illness
- To mediate alerting, used ESSENCE quality measure discharge diagnosis informative (DDI), the daily percentages of discharge diagnoses that are “informative”
 - Median DDI of 94% over all counties, picked those below 70% overall for analysis



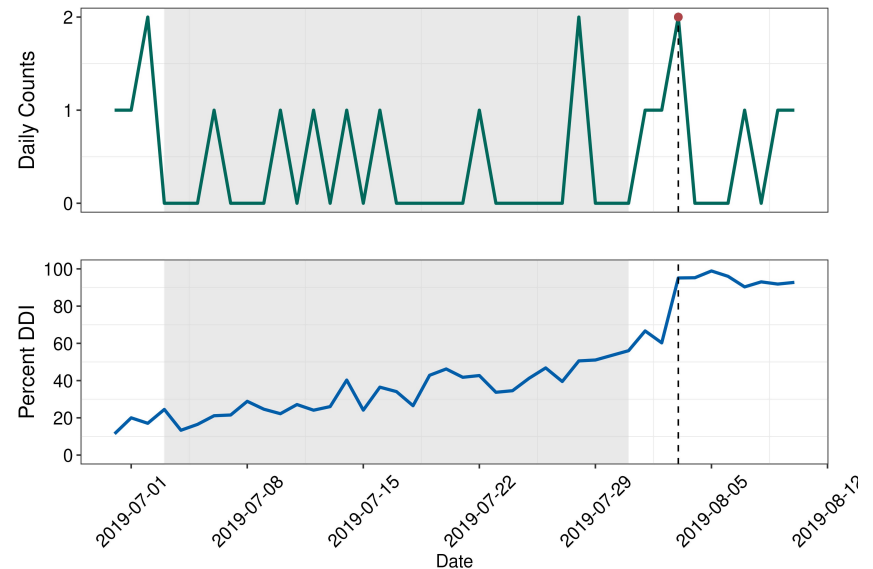
Examples of Alerts Due to Poor Data Quality

Clear artificial anomalies caused by large jumps in percent DDI towards end of 28-day baseline (row 2)

Baseline Counts and Percent DDI - Example County 1



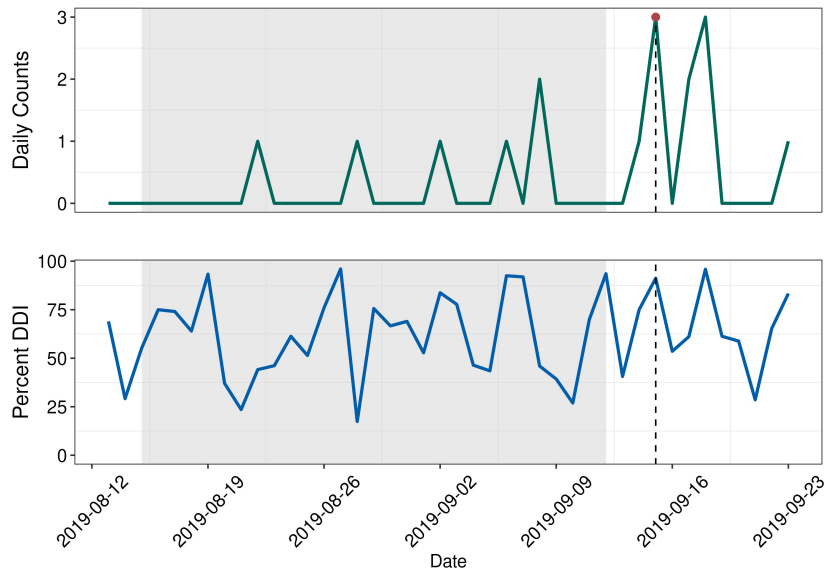
Baseline Counts and Percent DDI - Example County 2



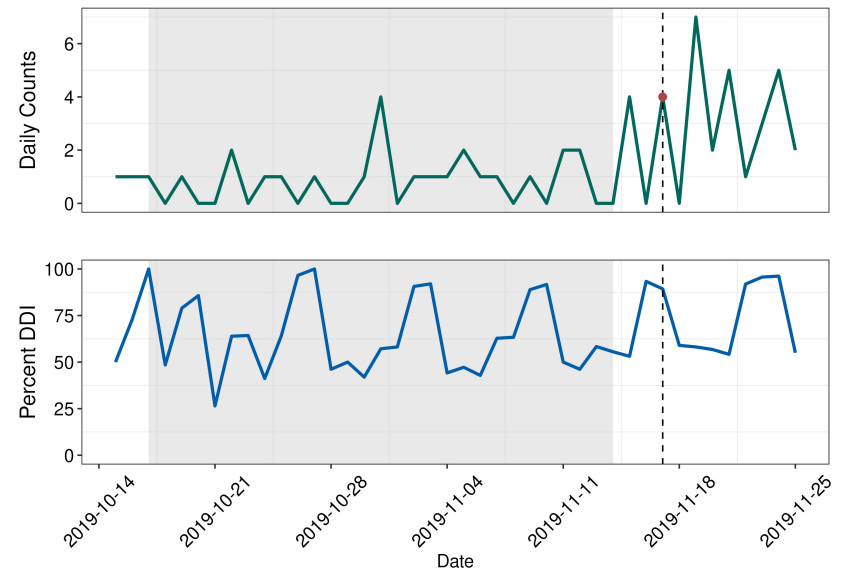
Examples of Alerts Due to Poor Data Quality

Unclear artificial anomalies caused by large jumps in percent DDI towards end of 28-day baseline – high variation in percent DDI over baseline (row 2)

Baseline Counts and Percent DDI - Example County 1



Baseline Counts and Percent DDI - Example County 2

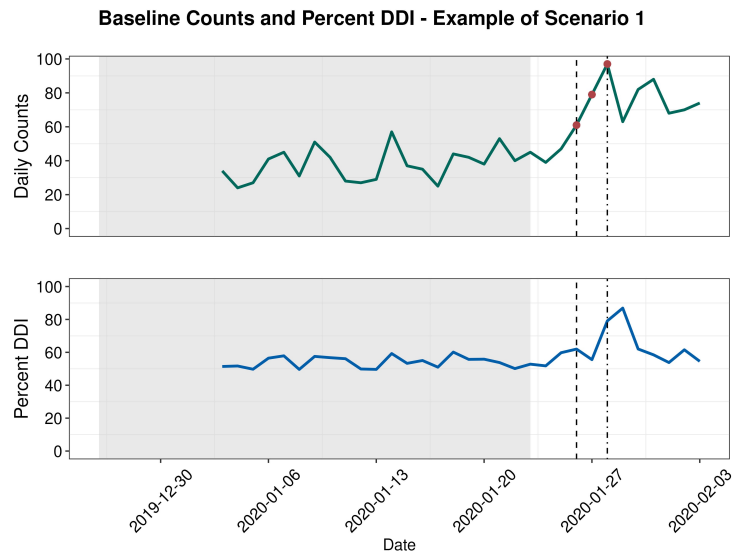


Quality Alert Filtering Approach: Principles Adopted

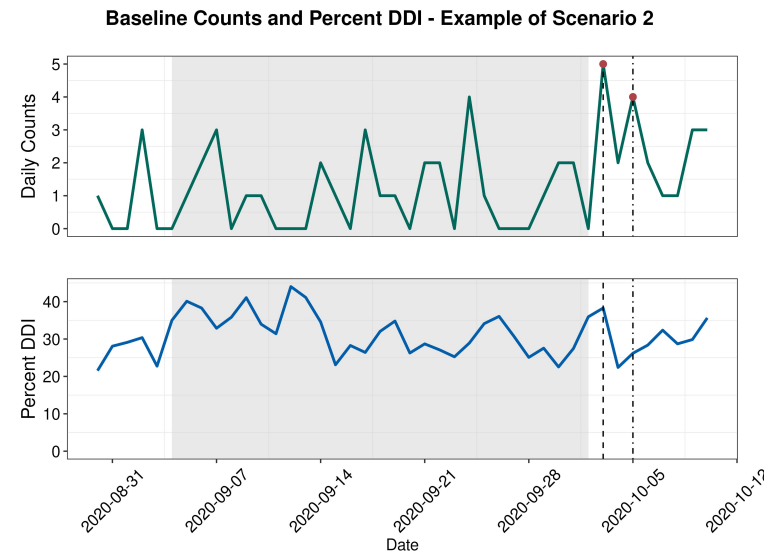
- For each algorithm anomaly, apply discriminator and either retain as alert, or drop
- There are no partial algorithm alerts
 - An alert is either accepted or rejected
- The only adjustment concept is to suppress alerts that seem to result from a data quality jump
 - Not to infer alerts when data quality has dropped

Quality Alert Filtering Approach: Principles Adopted – Exceptions

Scenario 1: In a string of consecutive alerts, if early ones are accepted, then later ones are also accepted even if data quality abruptly rises



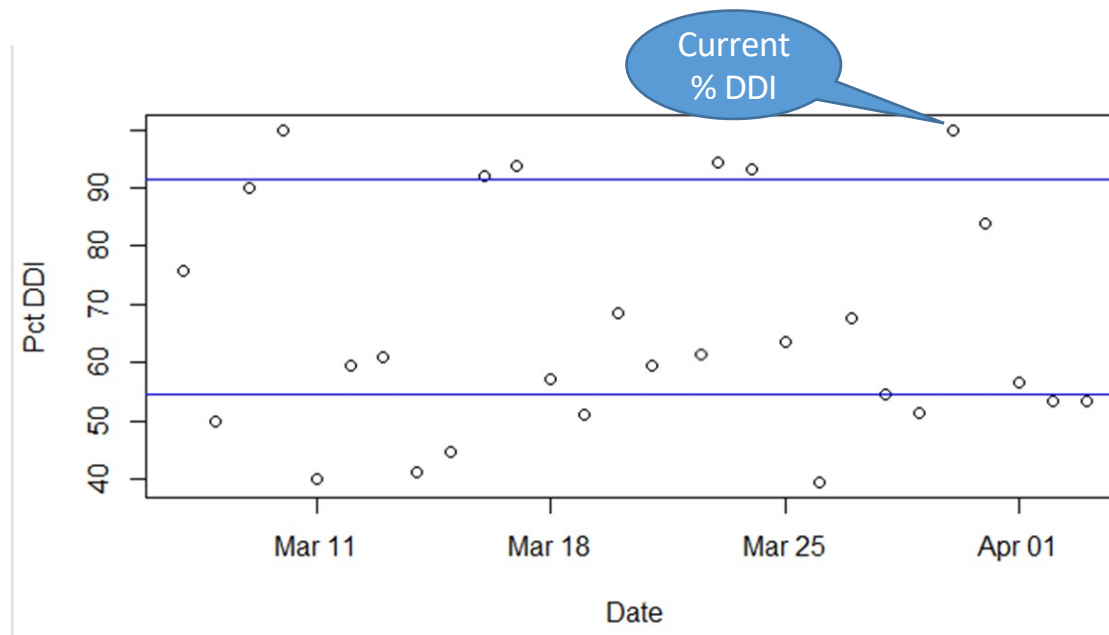
Scenario 2: However, if alerts are rejected because of a quality gap, and the alerts continue even after data quality falls to baseline levels, then later alerts are accepted



Adopted Approach

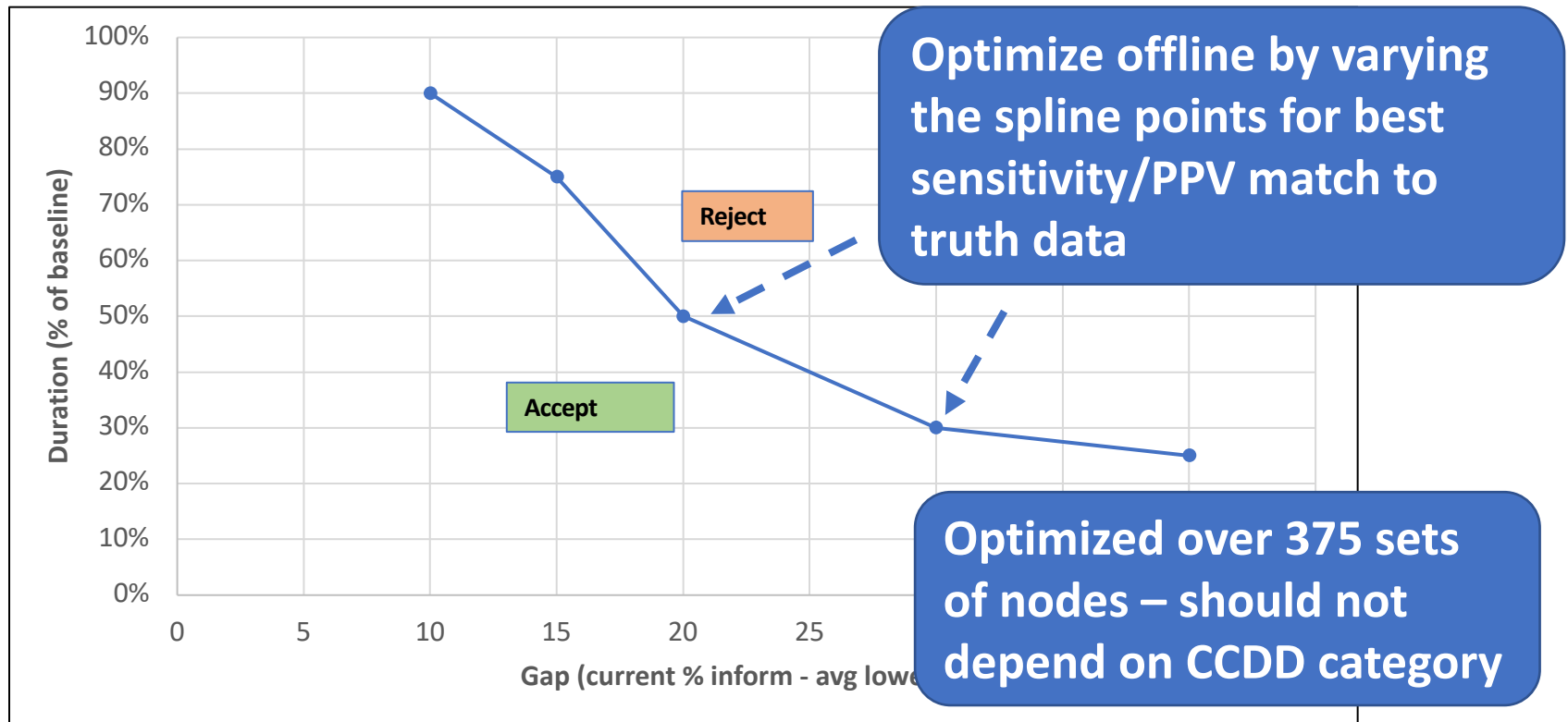
- Regard baseline as mixture of low-quality and high-quality data
 - μ_1 : Average percent discharge diagnosis informative for **low-quality** baseline
 - λ_1 : Ratio of baseline for **low-quality** data
 - μ_2 : Average percent discharge diagnosis informative for **high-quality** baseline
 - λ_2 : Ratio of baseline for **high-quality** data
 - Gap: = Current percent DDI - μ_1
- Decide whether to reject based on combination of Gap and low-quality ratio λ_1
- Is the data quality high enough for long enough to keep alert?

Calculation of Alert/Baseline Quality Gap



How large is the data quality gap? For how much of the baseline?

Use of Heuristic Accept/Reject Values to Calculate Smoothed Spline



Truth Data Derivation for Calibrating Discriminator

- Test data sets for choosing which alerts to reject because of jumps in data quality
 - Time series of counts, % DDI, and p-values for every region of low data quality (filtering to regions with overall % DDI < 70% **and** at least 2 alerts per month)

Data Set	Dates	Number of Regions	Number of Alerts
Influenza-like Illness (ILI) CCDD Category	Jan 1 st – Dec 31 st , 2019	131	1,671
	Jan 1 st – Dec 31 st , 2020	99	1,359

- Percentage of rejected alerts varied from 7% to 11%
 - “Right answer” depends on year, state of data quality

R Shiny App for Alert Evaluation

Summarized agreement over each data set based on manual review by both Michael Sheppard and Howard Burkom, using Shiny app developed in collaboration with Roseric Azondekon for annotation

Select fips code
Sample FIPS

Select alert number
4

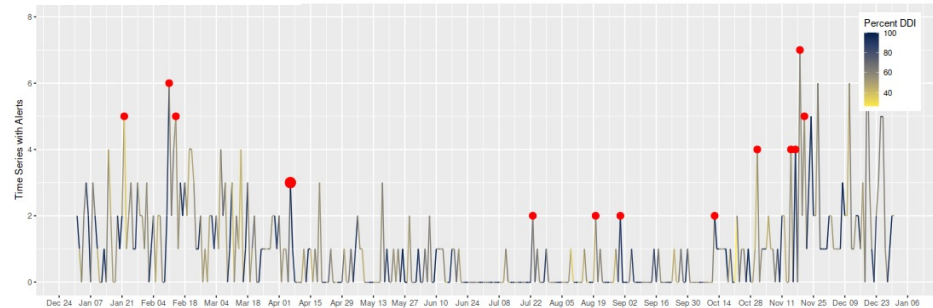
Save Changes Next Alert Next FIPS

Show 10 entries Search:

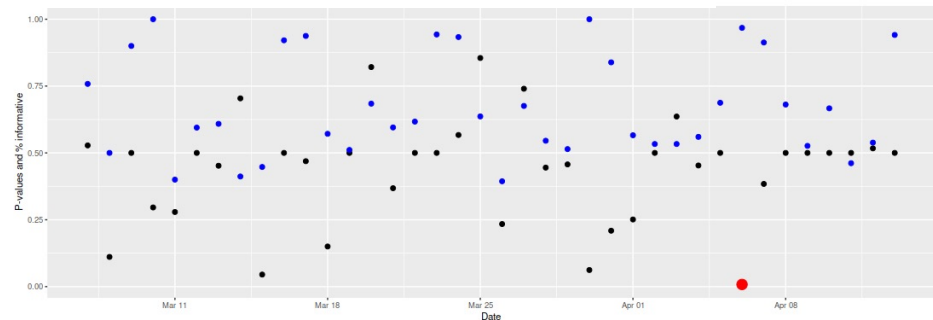
	date	fips	tsValue	pValue	pctInform	okAlertFlag
1	2019-01-01	Sample	0	0.5	0	
2	2019-01-02	Sample	0	0.5	0	
3	2019-01-03	Sample	0	0.5	0	
4	2019-01-04	Sample	0	0.5	0	
5	2019-01-05	Sample	0	0.5	0	
6	2019-01-06	Sample	0	0.5	0	
7	2019-01-07	Sample	0	0.5	0	

Showing 1 to 10 of 47,815 entries Previous 1 2 3 4 5 ... 4782 Next

Time Series for Sample FIPS



p-values (black, red for alerts) and % informative (blue) for Sample FIPS



Quality Alert Filter Performance Relative to Truth Data Sets

Data Set	Optimal Spline Values									Performance Metrics			
	1	5	10	15	20	30	40	50	90	Sensitivity	PPV	Accuracy	F-Measure
ILI 2019	1.00	0.95	0.95	0.80	0.60	0.35	0.30	0.25	0.25	93.8%	42.3%	94.3%	58.3%
ILI 2020	1.00	0.95	0.95	0.80	0.50	0.30	0.30	0.25	0.25	76.3%	55.8%	94.5%	64.4%



Only minor differences in optimal spline knots among test sets

Description of Recommended Algorithm

- For each alert:
 - Apply simpleMix algorithm to baseline time series (28 days) to derive μ_1 , μ_2 , λ_1 , and λ_2 , then quality gap
 - Modify gap for recency according to bubble-sort metric
 - Reduce gap if enough of more recent baseline data have high quality
 - Calculate threshold for gap from the smoothed spline for the maximum acceptable low-quality baseline percentage λ_1 (spline previously computed offline)
 - (e.g., “For what percentage of the baseline is a 25% gap acceptable”?)
 - Suppress alert if the gap exceeds the threshold

Component Calculations

- Algorithm `simpleMix()` (25 lines of base R code using only mean and variance calls)
- Offline optimization of spline coefficients
- Bubble-sort metric for recency of high-quality baseline data
- Interpolation of the current gap into the spline to get the acceptance threshold
- **Note:** Algorithm can be included at the end of any code implementing existing ESSENCE anomaly detection algorithms, given daily DDI percentages

Question and Answers

Thank you for your attention!



Contact Information

- **Howard Burkom:** Howard.Burkom@jhuapl.edu
- **Michael Sheppard:** oul2@cdc.gov
- **Roseric Azondekon:** poj2@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

National Syndromic Surveillance Program

Community of Practice

Resources that Advance the Science and
Practice of Syndromic Surveillance

2021

Syndromic Surveillance Symposium

Week of November 15, 2021



KAHUINA
CONSULTING