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

KAHUINA CONSULTING

CSTE

Council of State and Territorial Epidemiologists

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

CSTE

KAHUINA CONSULTING

Applying supervised machine learning to query emergency medical service databases for cases of opioid overdose

NSSP Symposium 2021

Hannah P. Cowley, BA
Hannah.Cowley@jhuapl.edu

Robyn Ellis, MPH (Oregon Health Authority)
Breanna Swan, PhD (JHU/APL)
Howard Burkom, PhD (JHU/APL)

Conflict of Interest Statement

I declare that I have no conflicts of interest.

Identifying opioid overdose cases from the EMS record is difficult, but important.

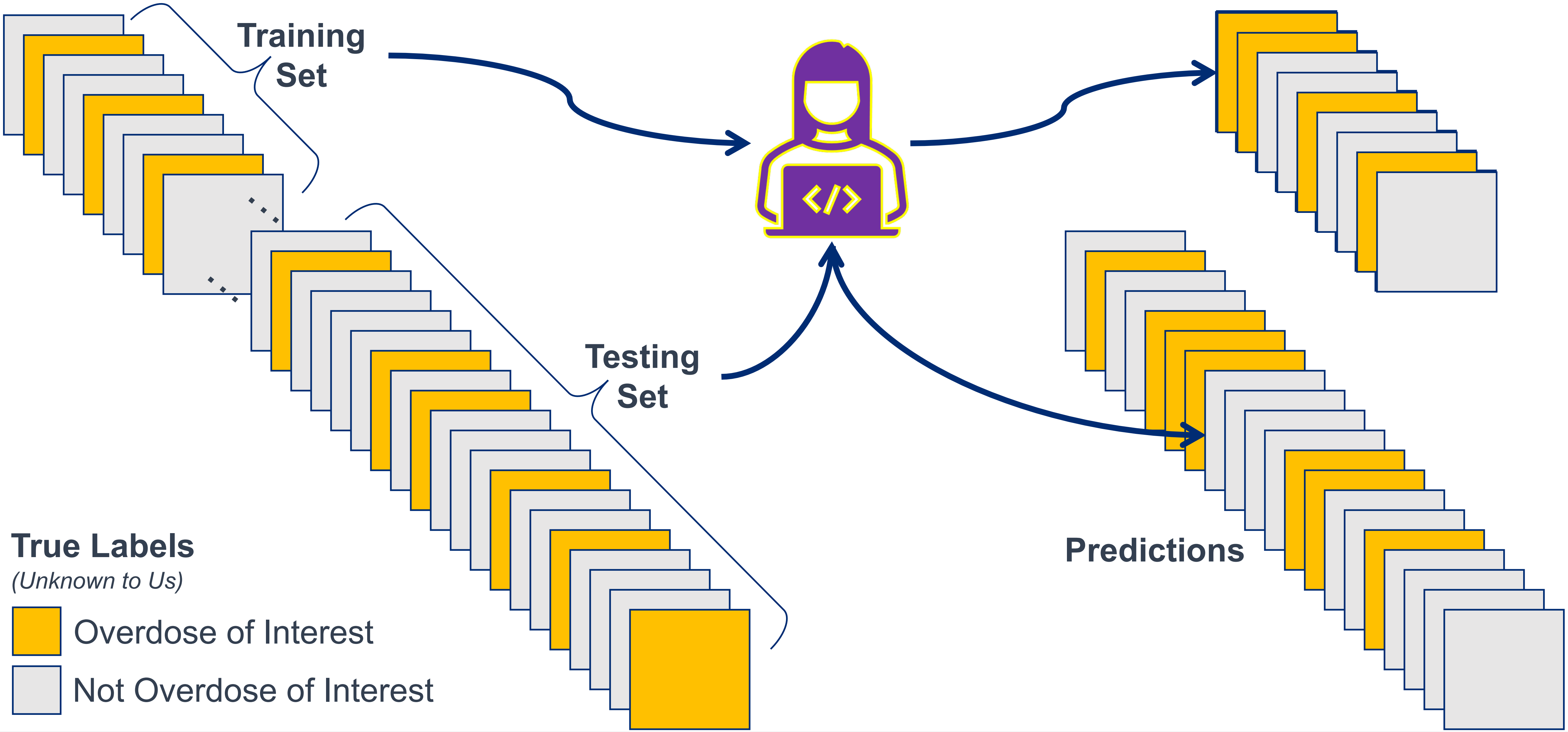
Image sources:
https://www.njpies.org/product/poison-help-magnets-english/
https://www.logolynx.com/topic/ems
https://www.ohsu.edu/health/ohsu-and-doernbecher-emergency-care

Identifying specific types of opioid overdoses can be even more challenging.

Overdoses of Interest: *Unintentional*

Exclude suicide attempts, medication errors

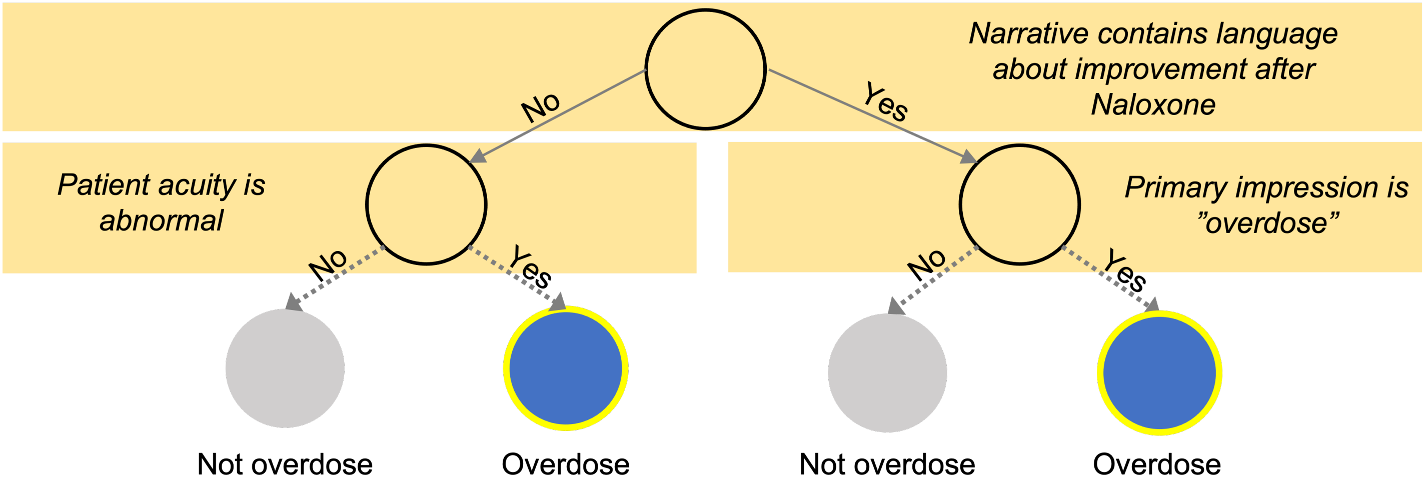
Introduction to Supervised Machine Learning



The Approach

- *Data:* 1,123,666 total EMS records from 1/1/2018 – 9/9/2020
- *Training set*: 1500 EMS records, annotated by our team (2 annotators per record)
- *Model:* A random forest classifier
  - A collection of 200 decision trees that each predict “overdose” versus “not overdose”
  - Set a threshold for the proportion of trees that must declare “overdose” for the record to be predicted as such
    - E.g., “Majority rules” – a voting threshold of 0.5

A sample decision tree



Feature Extraction and Encoding

- In order to use a machine learning algorithm, we must convert relevant fields to numeric values

Training and Testing the Model

- 5-fold cross-validation was used
- Multiple voting thresholds were investigated to see which one gave the best balance of precision (positive predictive value) and recall (sensitivity)

Voting Threshold	Precision (PPV)				Recall (Sensitivity)			
	Mean	*Std*	*Minimum*	*Maximum*	*Mean*	*Std*	*Minimum*	*Maximum*
0.25	**0.73**	0.03	0.70	0.76	**0.90**	0.05	0.84	0.97
0.30	**0.77**	0.02	0.74	0.80	**0.89**	0.06	0.82	0.97
0.35	**0.80**	0.03	0.78	0.84	**0.87**	0.06	0.81	0.95
0.40	**0.83**	0.05	0.78	0.88	**0.85**	0.05	0.79	0.94
0.45	**0.85**	0.05	0.81	0.91	**0.82**	0.05	0.76	0.90
0.50	**0.87**	0.05	0.80	0.93	**0.79**	0.07	0.71	0.89

Can reach 90% recall at the expense of precision

Can maximize precision at the expense of recall

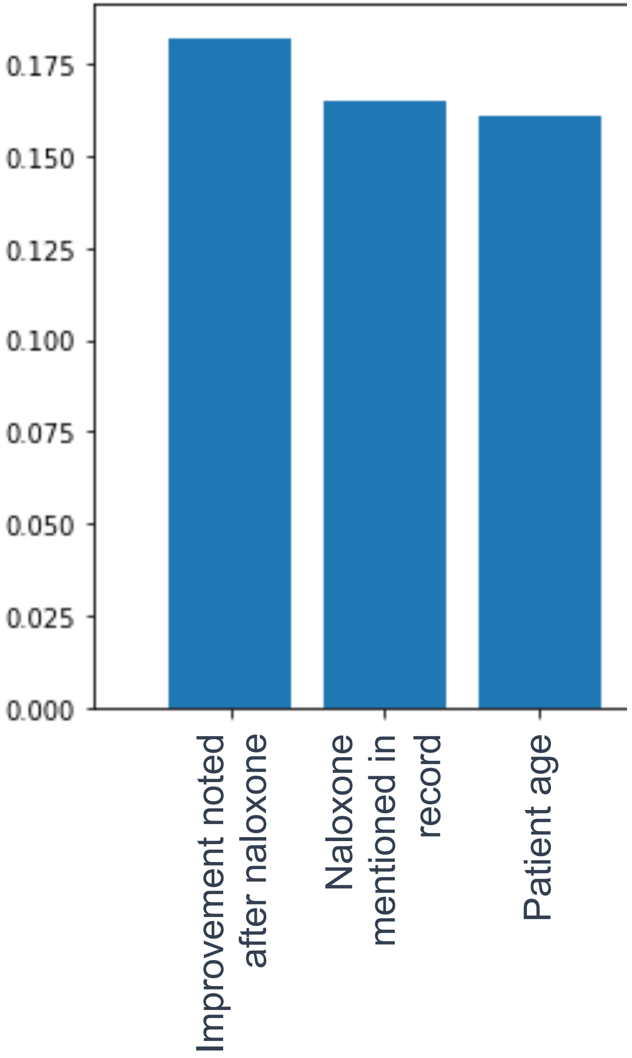
Our goal was to capture the majority of overdoses of interest, therefore we value high sensitivity.

Interpreting The Model

Importance

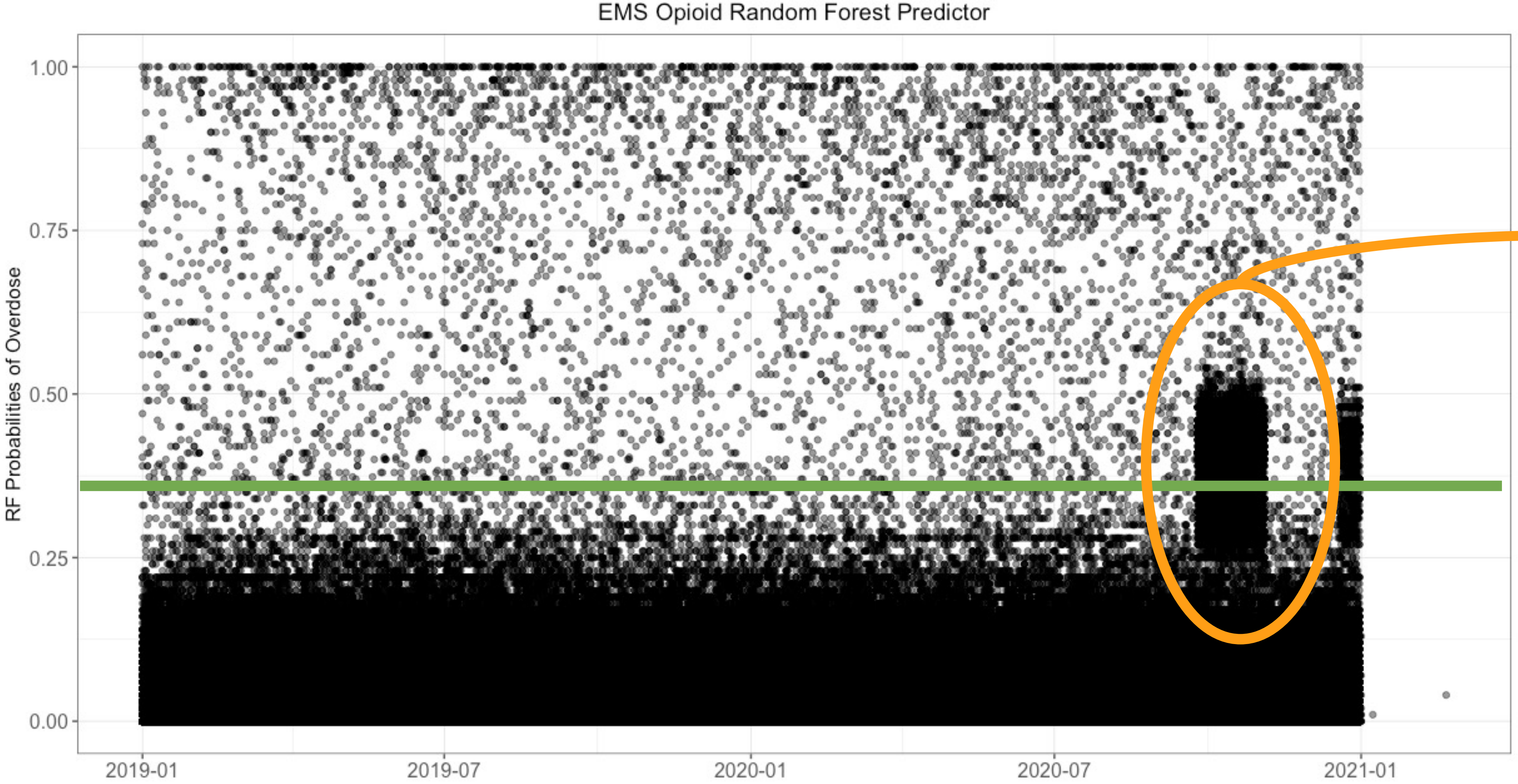
How many nodes in the tree are split by this feature, weighted by importance of the node in the tree

Feature importance of top 10 features



Operationalization

- 13,877 records of the 1 million+ were predicted to be "overdoses"

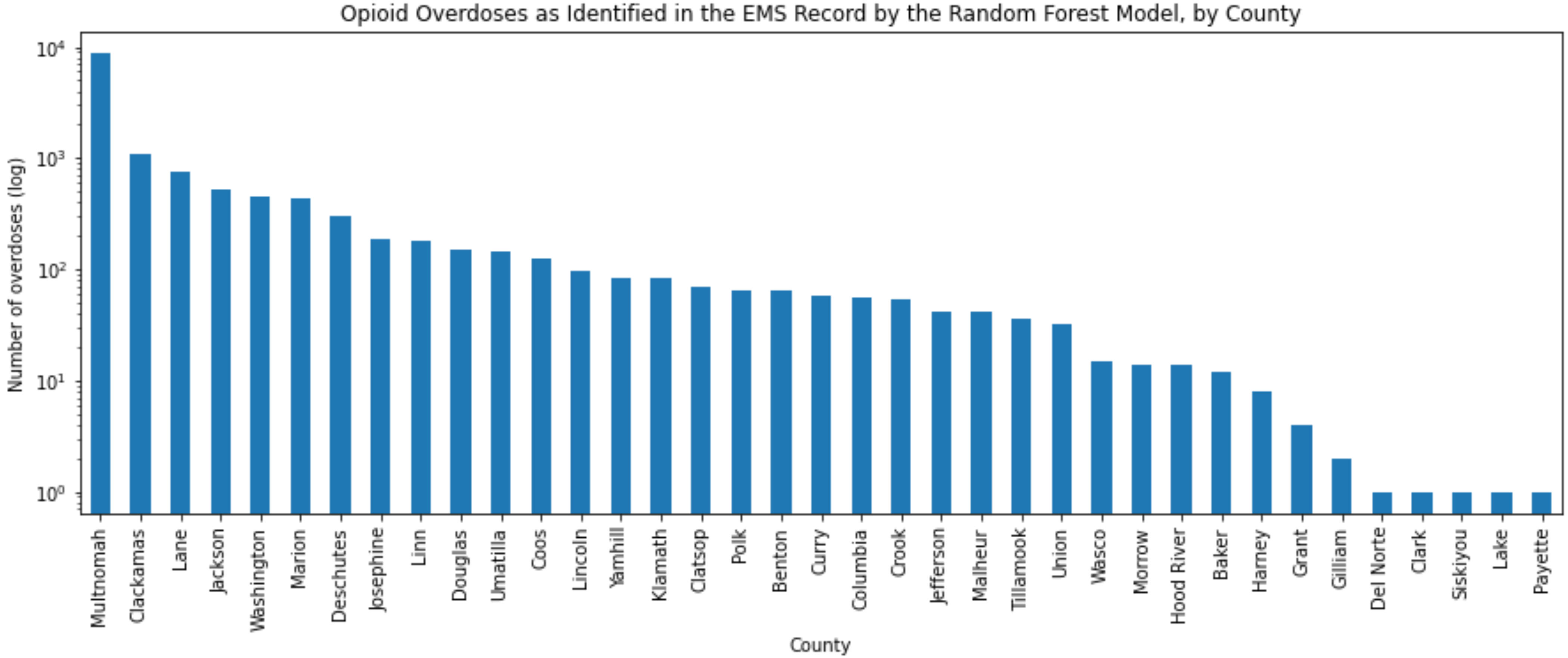


One challenge in operationalizing: When the narrative is missing, the model tends to vote in favor of "overdose"

Approx. random forest voting threshold

Operationalization: What Now?

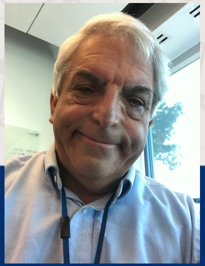
- Records identified by the model can be used directly for public health surveillance, or in other modeling efforts



Limitations and Challenges

- Challenges:
  - Deciding which cases of opioid overdose are of interest
    - Here, we focus on accidental overdoses from recreational use
      - We *do not* focus on suicidal ideation or accidental prescribed medication overdose
  - Preprocessing of data and standardization
    - Pulling out medications of interest from the medication and narrative fields
    - Deciding how to group features and encoding numerically
  - Models are only as good as the data they are provided
    - For instance, our Classifier yielded a false spike during September 25th, 2020 to November 4th, 2020 – the result of missing narrative fields from one major agency.
    - Without realizing it, we had trained the model to rely heavily on these narrative fields
- Limitations:
  - Generalizing the exact implementation across geographic areas, time
  - Validating the accuracy of modeling is imperfect
  - Accessing these data and standardization across states
    - 33 states use EMS data for surveillance

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY



Michael Sheppard, MS – CDC &

Howard Burkom, PhD - Johns Hopkins University Applied Physics Laboratory

Centers for Disease Control and Prevention

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

NSSP National Syndromic Surveillance Program BioSense Platform

November 18, 2021

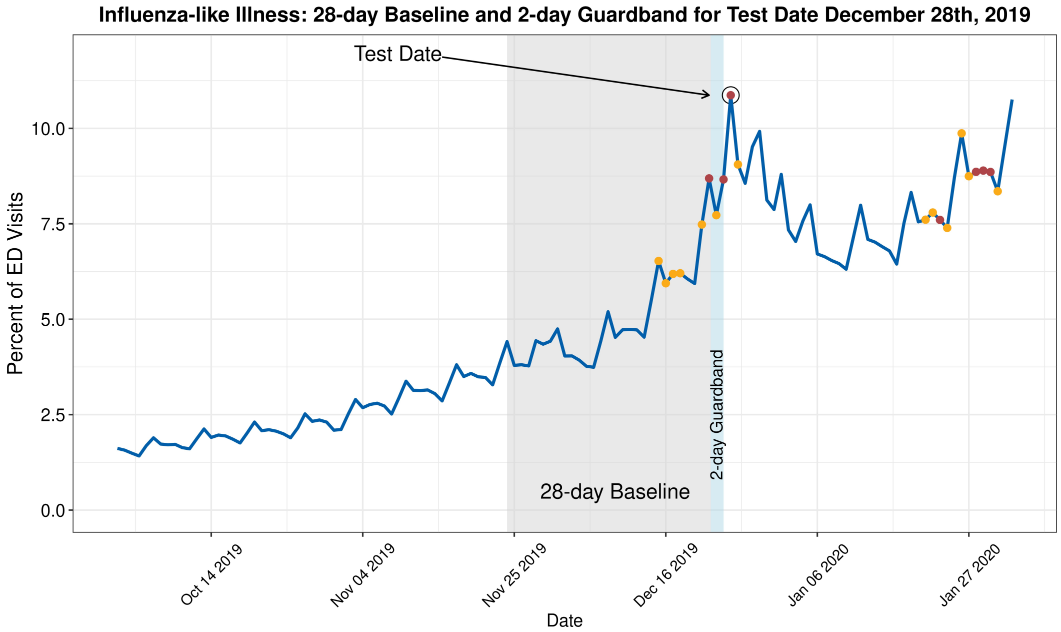
Issues of General Concern in Surveillance

- 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

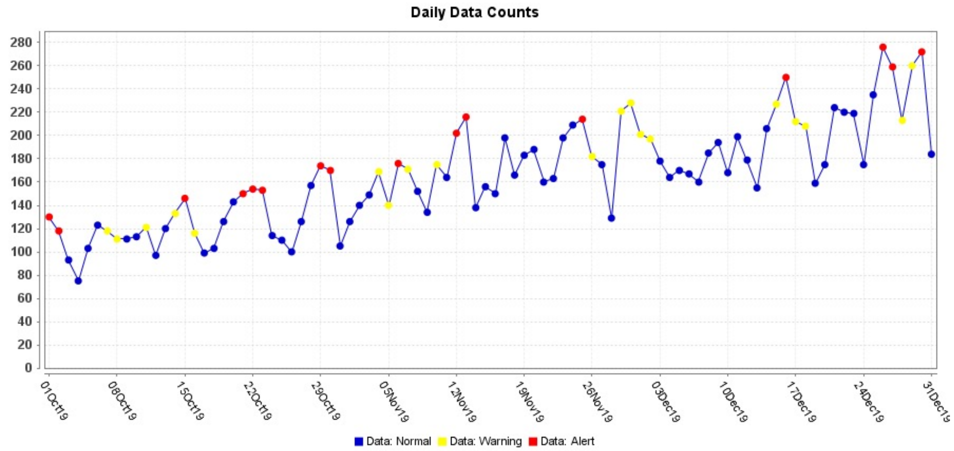
- Expectations need to be updated in many applications
- In ESSENCE systems, we use a sliding 28-day baseline with a 2-day guardband to separate the baseline from the test date



ESSENCE Datasets Used

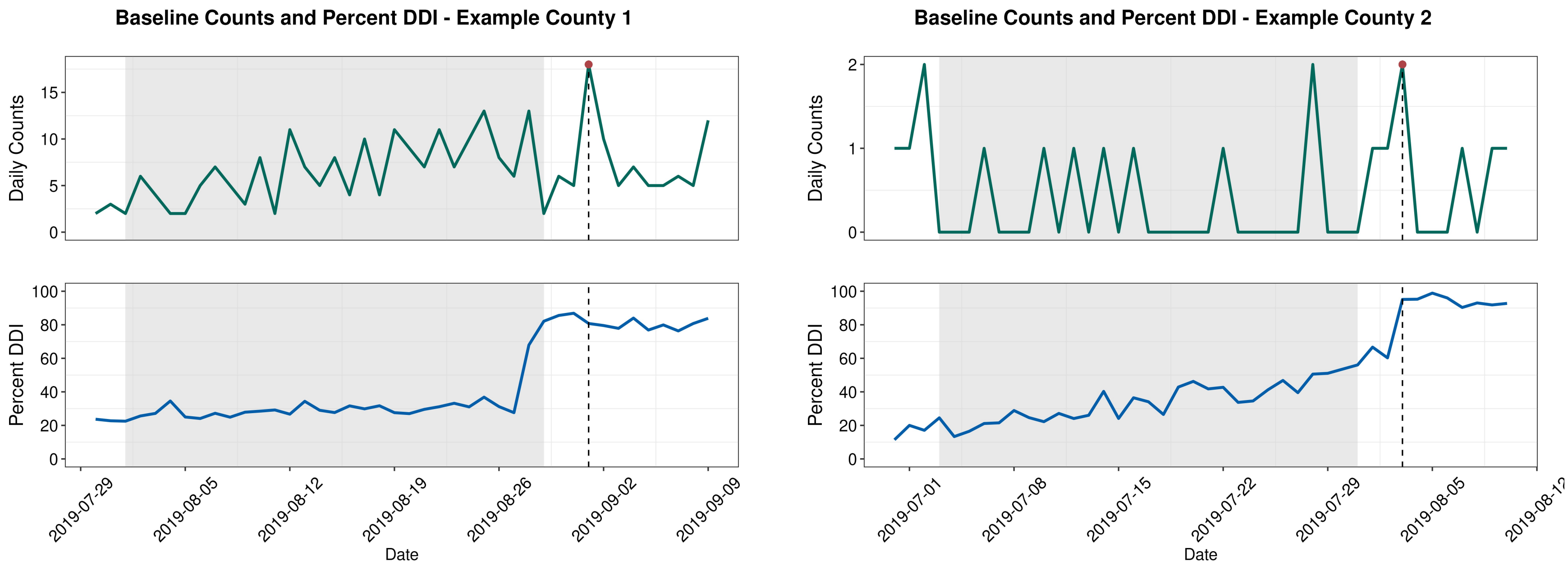
From counties with clearest data quality problems:

- 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



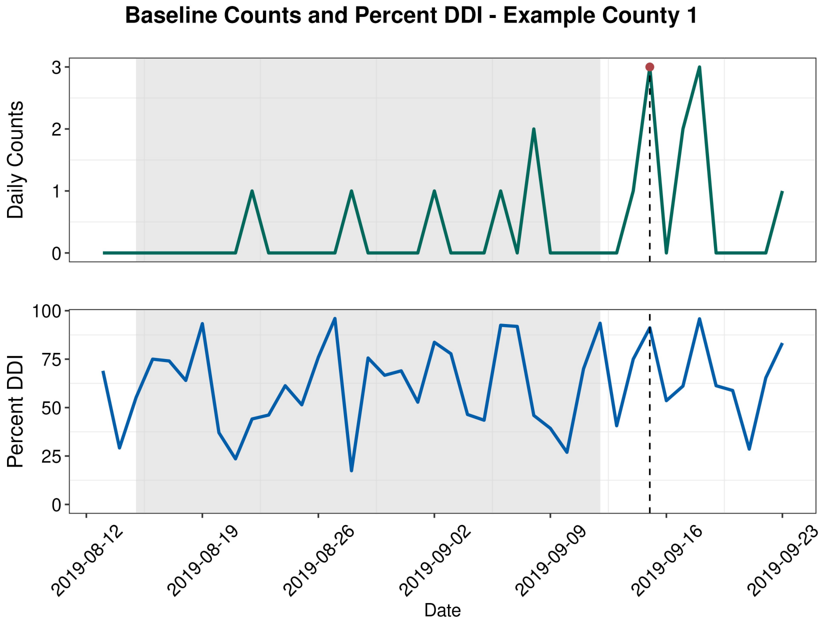
Examples of Alerts Due to Poor Data Quality

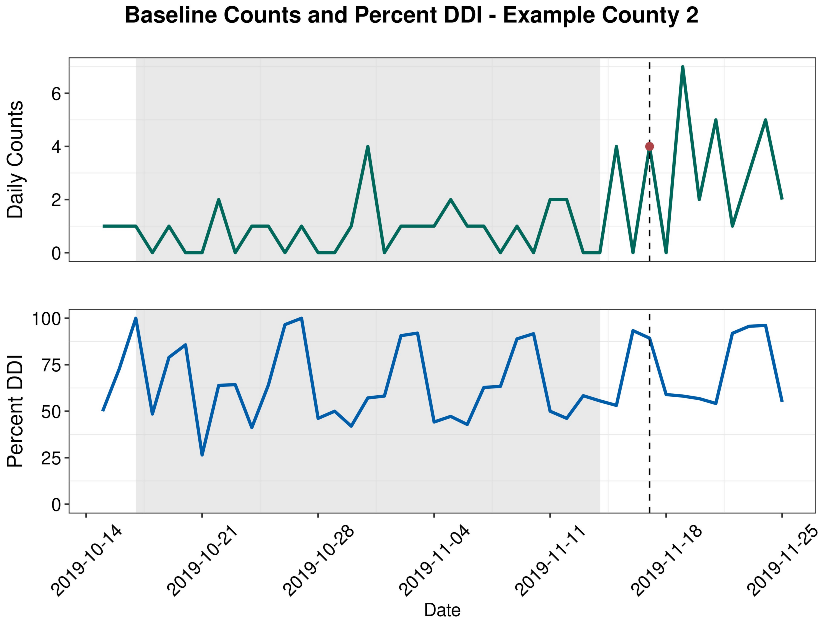
<u>Clear</u> artificial anomalies caused by large jumps in percent DDI towards end of 28-day baseline (row 2)



Examples of Alerts Due to Poor Data Quality

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



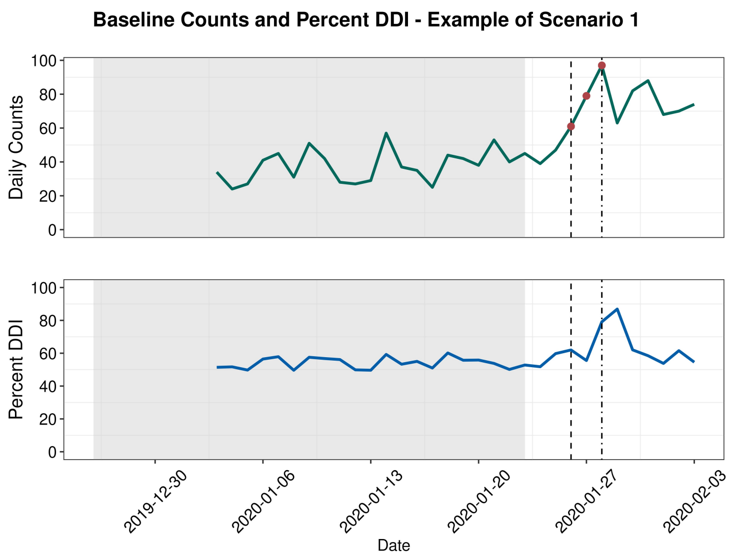


Quality Alert Filtering Approach: Principles Adopted

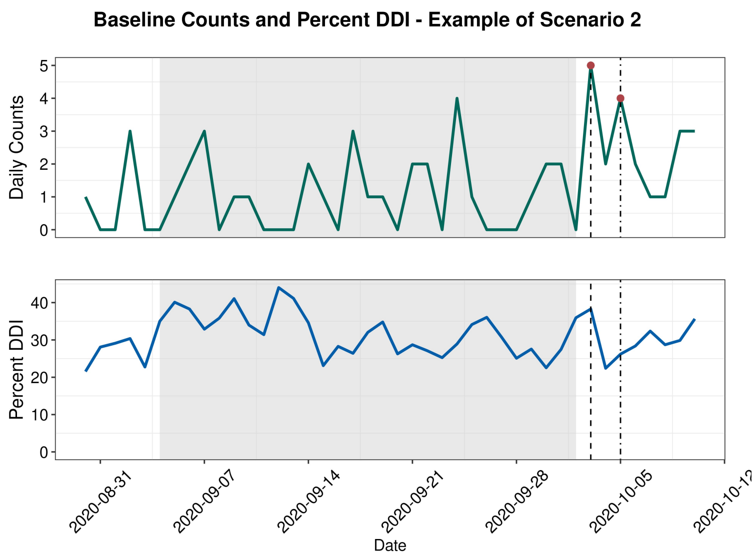
- For each algorithm anomaly, apply discriminator and either retain as alert, or drop
- <u>There are no partial algorithm alerts</u>
  - An alert is either accepted or rejected
- The only adjustment concept is to suppress alerts that seem to result from a data quality <u>jump</u>
  - Not to <u>infer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



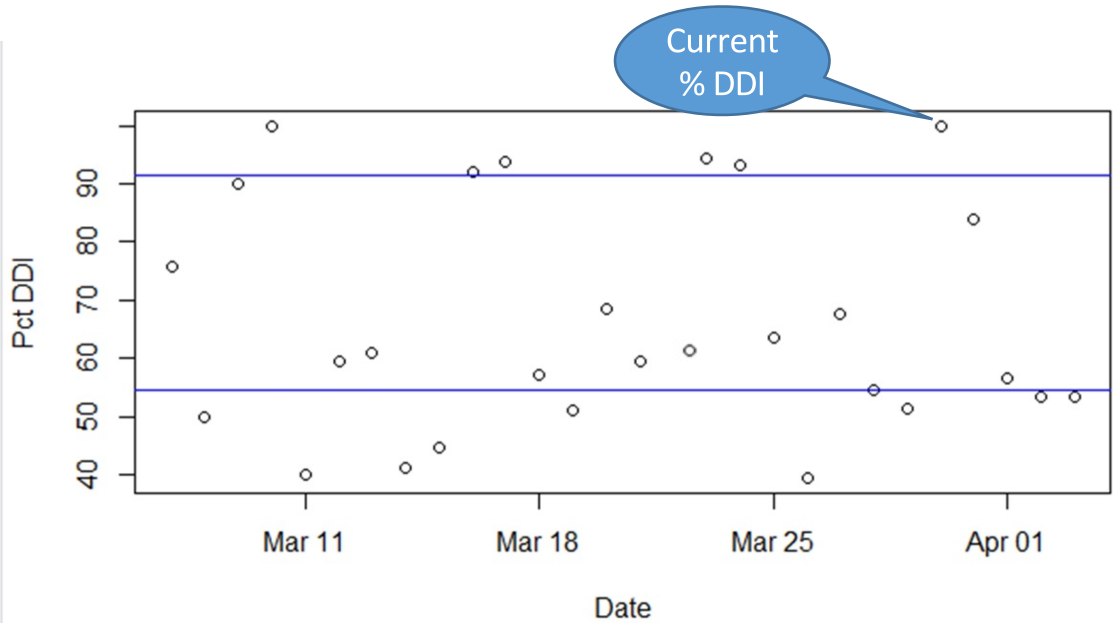
Scenario 2: However, if alerts are rejected because of a quality gap, and the alerts continue even after data quality falls to baseline levels, then later alerts are accepted



Adopted Approach

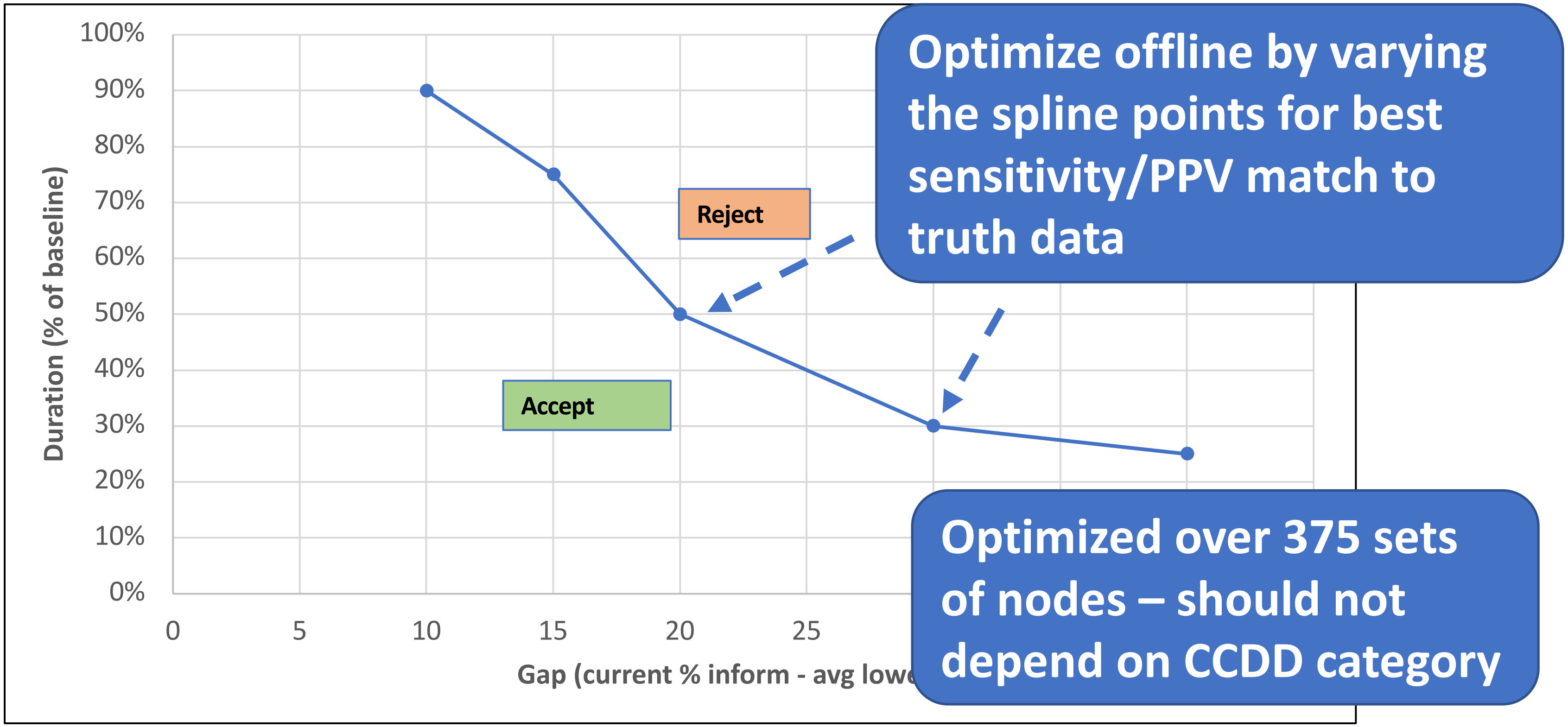
- Regard baseline as mixture of low-quality and high-quality data
  - μ_1: Average percent discharge diagnosis informative for **low-quality** baseline
  - λ_1: Ratio of baseline for **low-quality** data
  - μ_2: Average percent discharge diagnosis informative for **high-quality** baseline
  - λ_2: Ratio 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



<u>How large is the data quality gap? For how much of the baseline?</u>

Use of Heuristic Accept/Reject Values to Calculate Smoothed Spline



Truth Data Derivation for Calibrating Discriminator

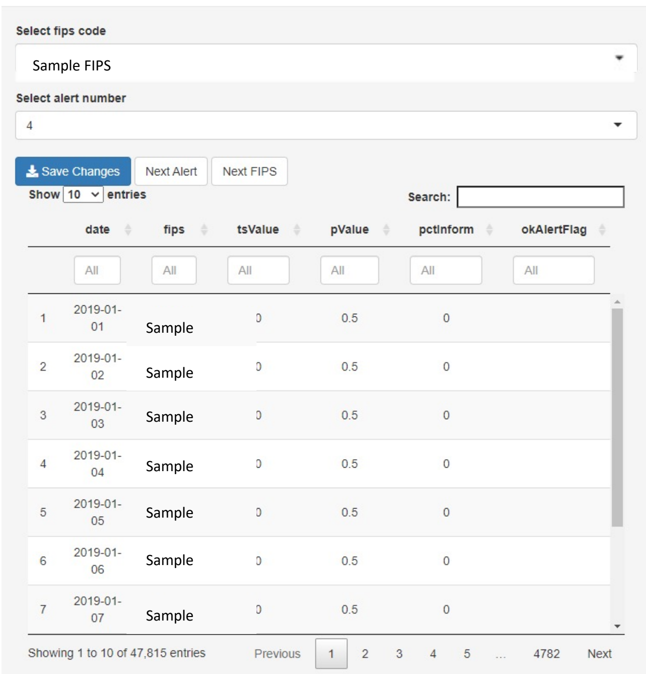
- Test data sets for choosing which alerts to reject because of jumps in data quality
  - Time series of counts, % DDI, and p-values for every region of low data quality (filtering to regions with overall % DDI < 70% **and** at least 2 alerts per month

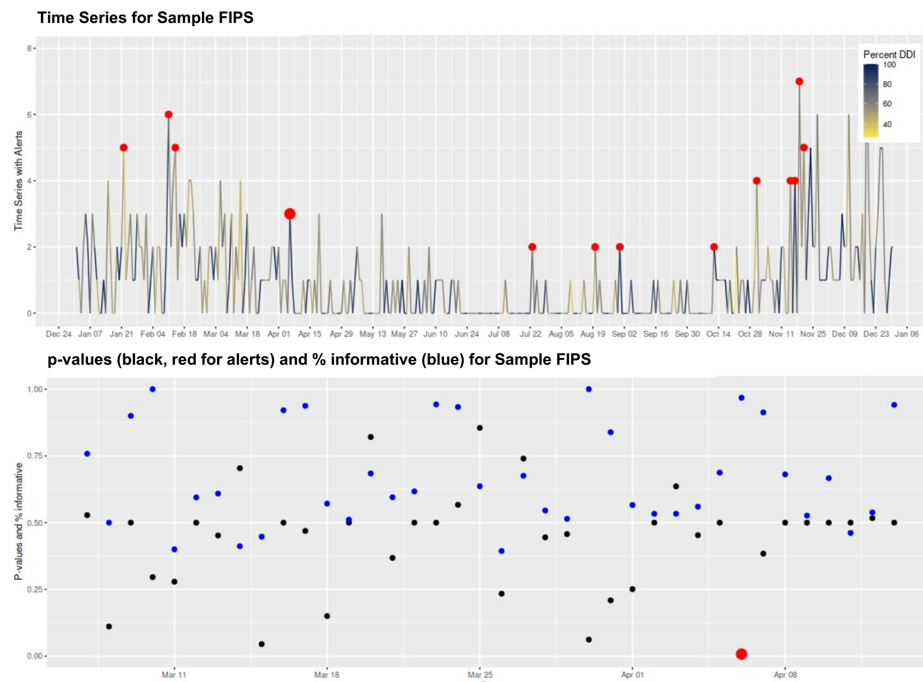
Data Set	Dates	Number of Regions	Number of Alerts
Influenza-like Illness (ILI) CCDD Category	Jan 1st – Dec 31st , 2019	131	1,671
	Jan 1st – Dec 31st, 2020	99	1,359

- 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





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

KAHUINA

CONSULTING