# **Ohio COVID-19 Surveillance and Screening-Shiny App Documentation**

# Graphical Display

# Contributors: Yuhan Pan, Mikkel Quam, Grzegorz Rempala, Patrick Schnell & Harley Vossler Contact: pan.387@buckeyemail.edu

Last updated 2020-06-09.

## Data structure

The primary data elements are the new case counts for each county on each day, arranged in the following columns:

1. Date (YYYY-MM-DD formatting, ~from 2020-03-01)
2. Count or Daily\_new (positive integer, ~0)

In addition, the following columns denote larger geographical areas:

1. County (unique character string, 88 in Ohio)
2. Region (unique character string, #1-8)
3. Zone (unique character string, #1-3)
4. State (unique character string, OH)

Within the application, areas are nested geographically such that each county belongs to exactly one region, each region to one zone, and each zone to one state. The definitions of the regions and zones are for predefined and in routine use by governmental decision makers.

A close up of a map

Description automatically generated

Finally, a final column indicates the estimated population of the county, according to the 2014–2018 5-year American Community Survey (ACS):

1. Population

These estimates are pulled from the ACS API via the tidycensus package in R.

Variables 1 and 2 will change as regularly as the data is updated (usually daily) while variables 3-6 are currently designed to be stable attributes of the counties. Data is obtained from The State of Ohio COVID-19 Cases: Ohio Department of Health – COVID-19 Dashboard.

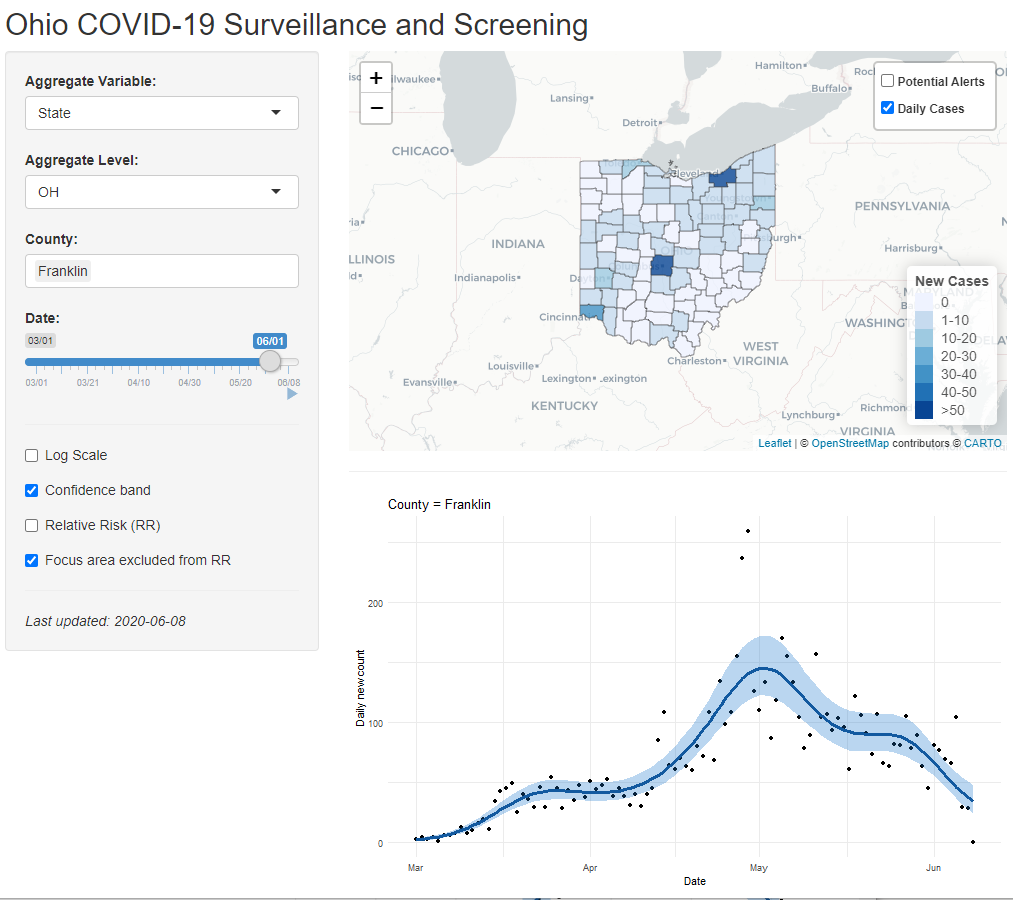
## Interactive Dashboard-Shiny Application User Interface

Maps, alerts, and analytical graphics on the right are generated according to user selections within the R Shiny application.

Maps and Alerts (upper right):

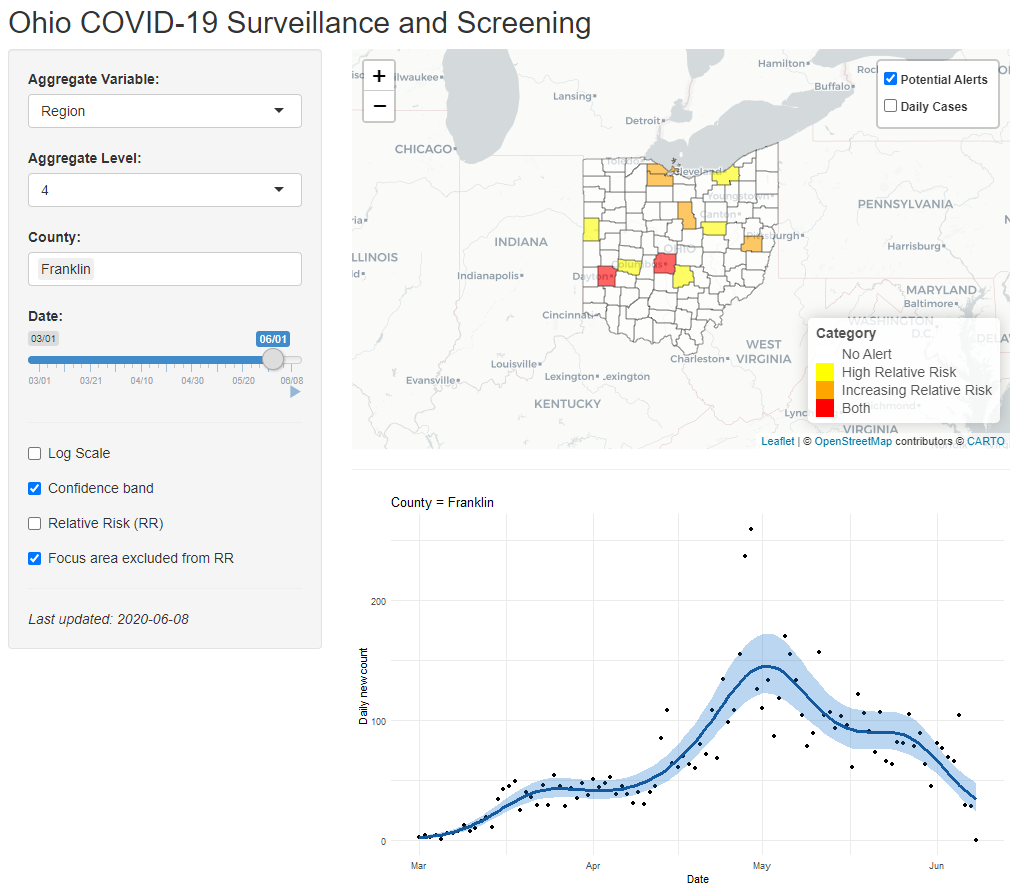
Within the R Shiny Application, the interactive base mapping is constructed using [Leaflet](https://rstudio.github.io/leaflet/). Users may zoom in and/or out to view an interactive base map of major road infrastructure for geographical orientation.

Daily Cases on June 1st



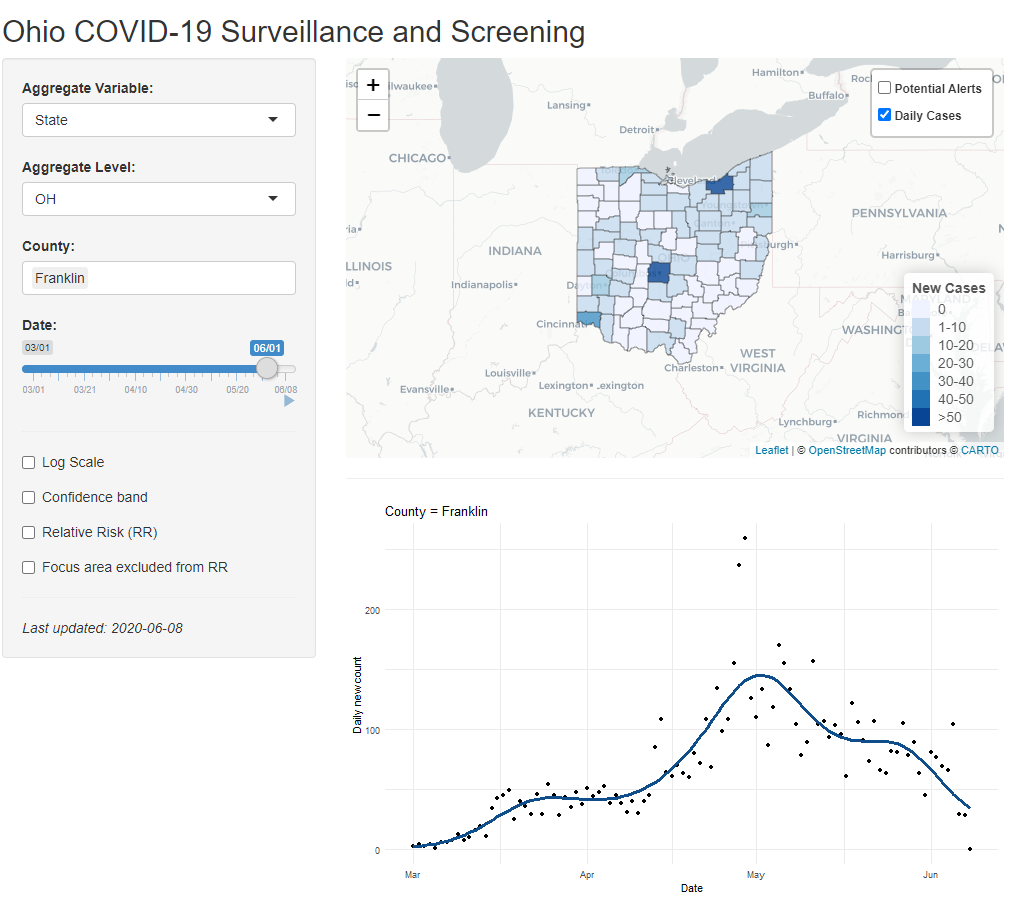
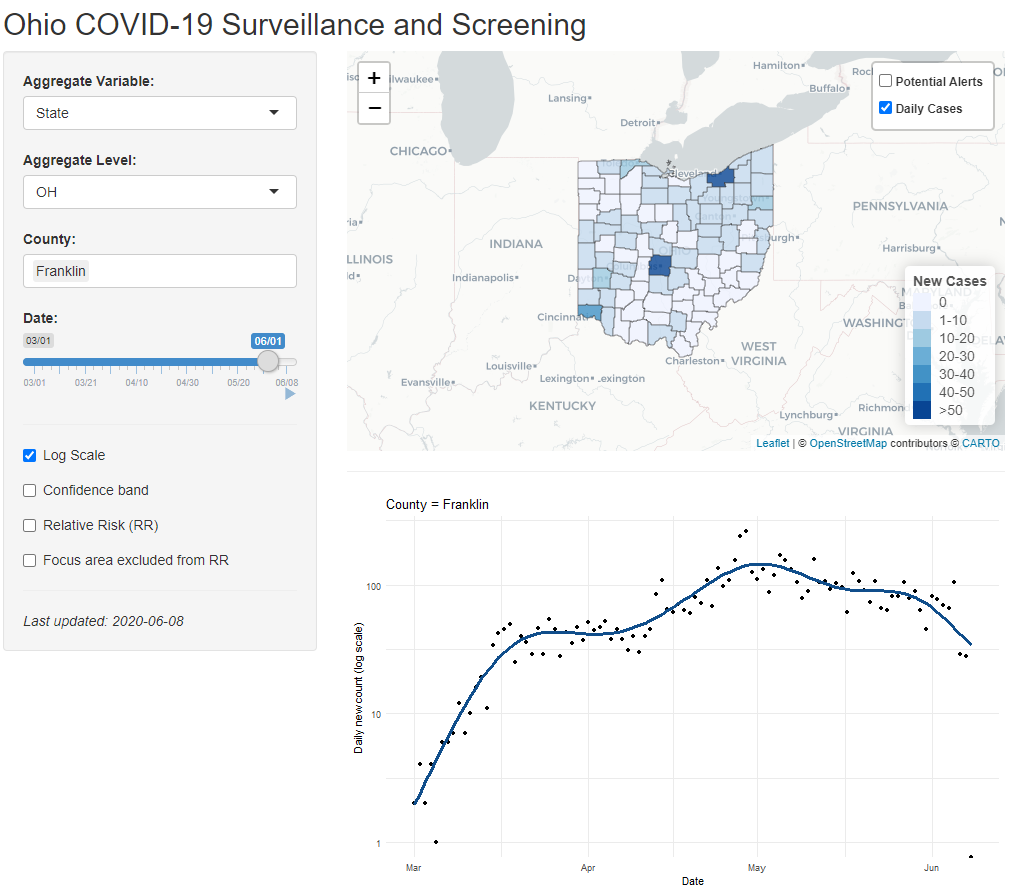
Users may determine to map layers containing information on alert screening and/or new daily case incidence according to checking boxes for layer in the upper right corner of the mapping pane. The current default maps new daily cases within the chosen county(ies) and alert screening (counties within compared to the entire state) for the most recent date in the dataset. Within a chosen county, users may also click on a specific county in the map to see the full number of new cases along with corresponding county, zone number and region number.

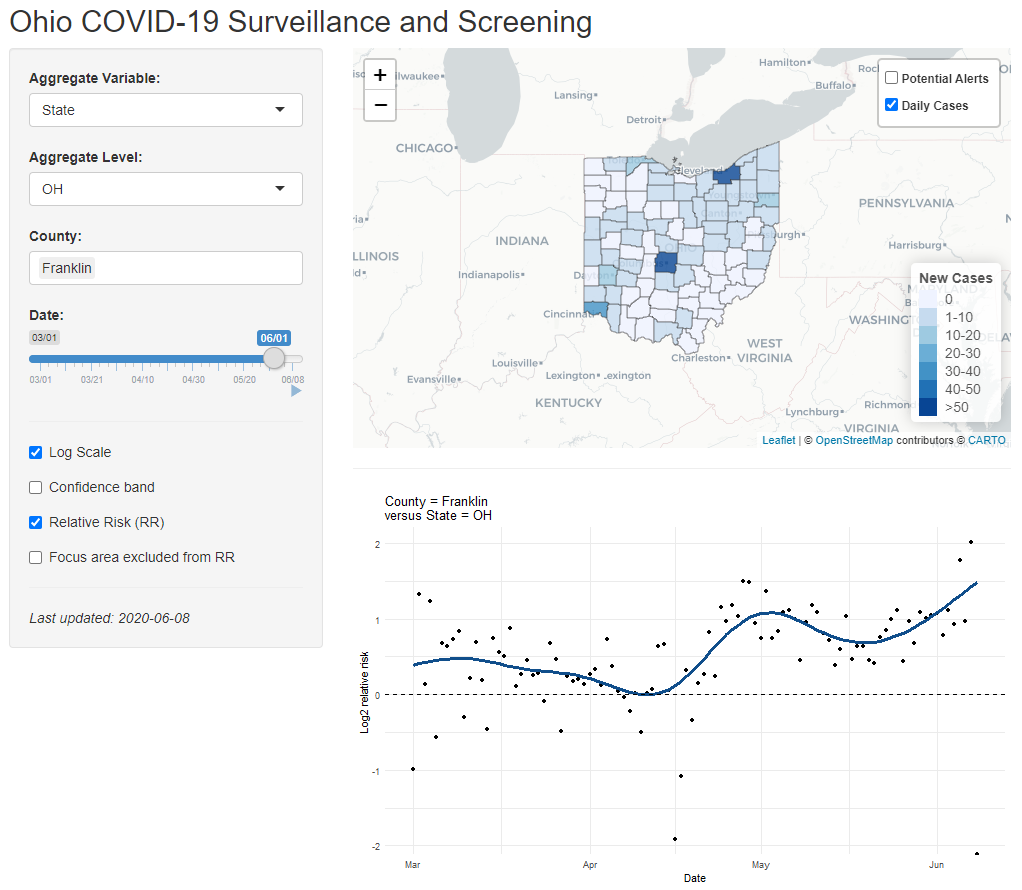
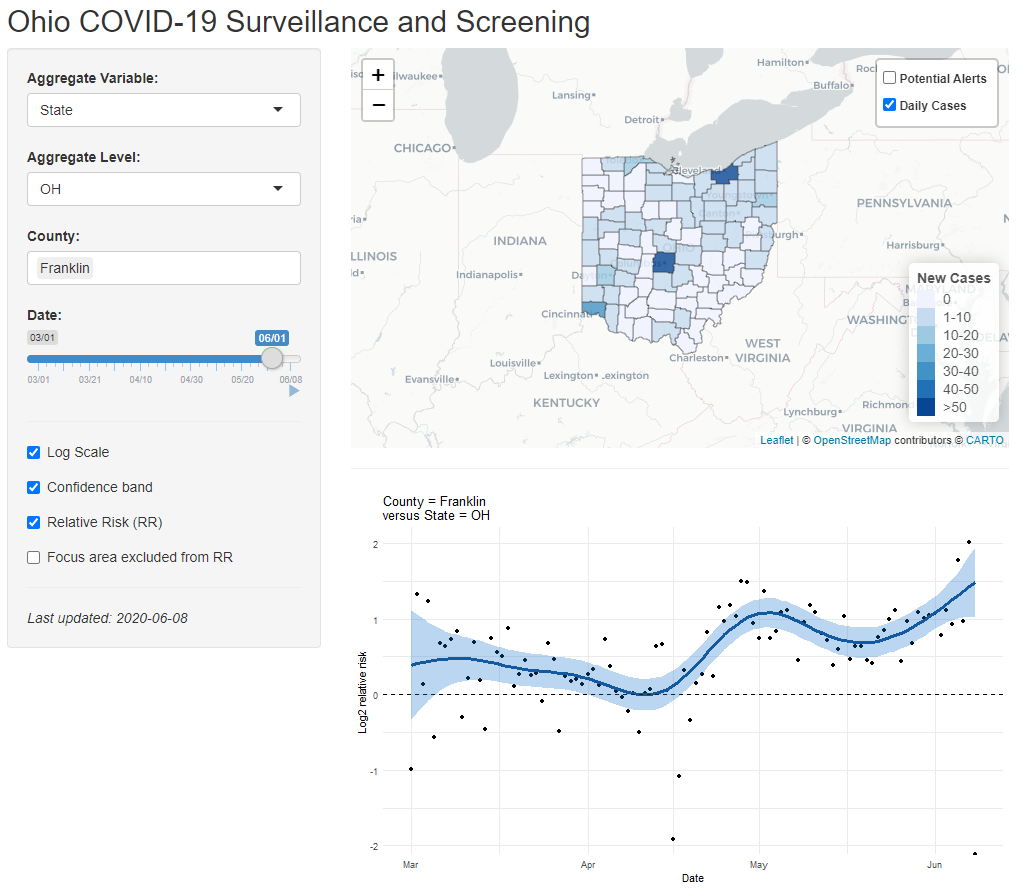
Potential Alerts, Daily Cases, and Further Inspection (Central, Ohio)

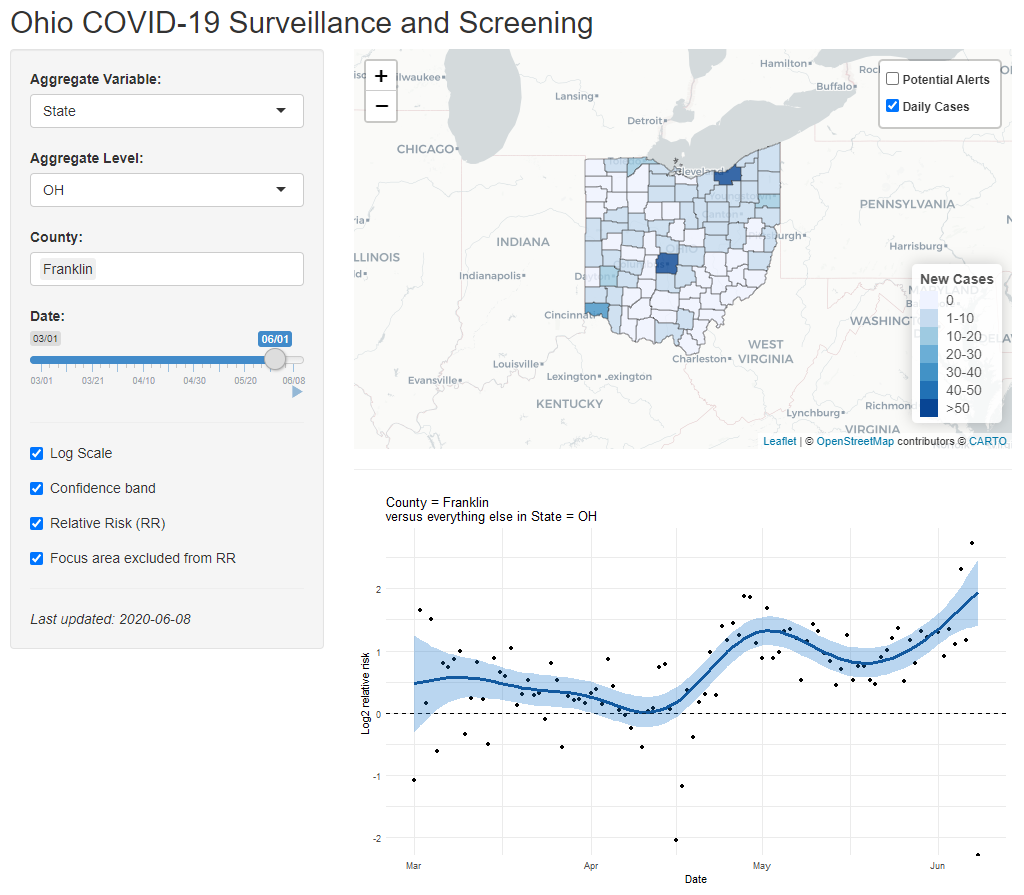
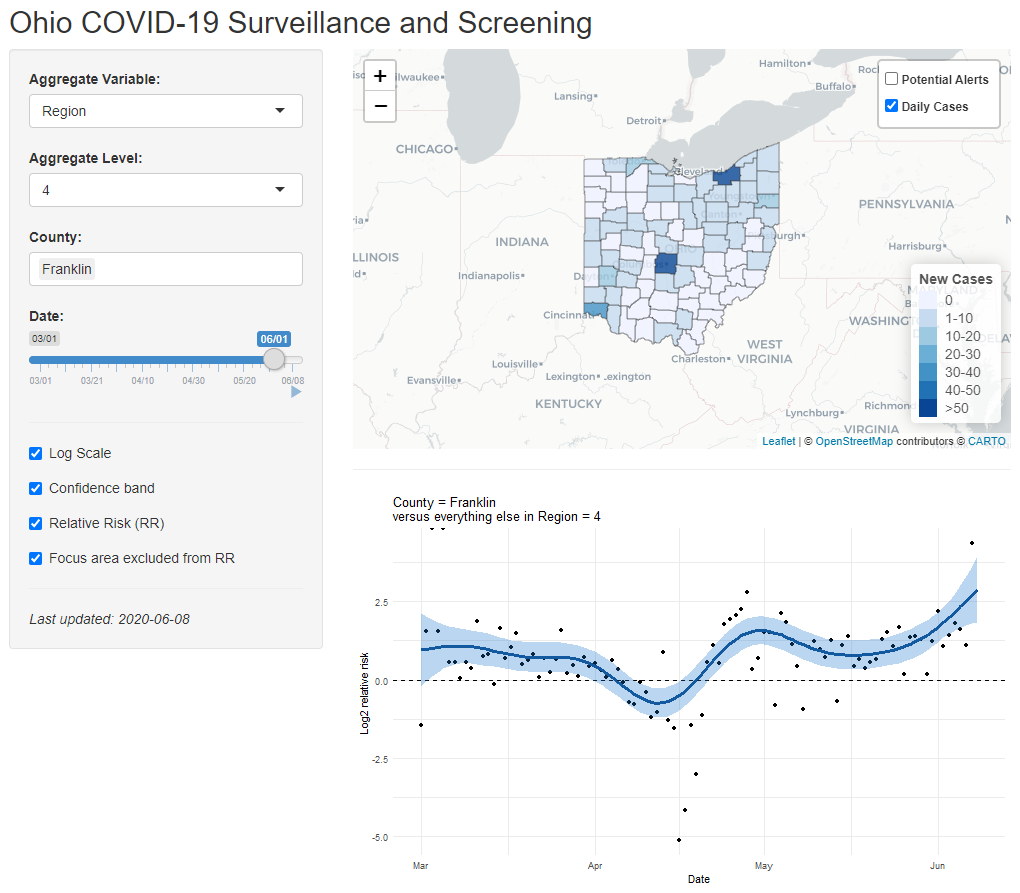


Even a few cases can cause an alert, if very few cases are being/have been reported in a given area. This may or may not be a real public health problem but merits some further inspection as to why the relative risk is increasing and/or high are the currently screen timepoint. Even though cases daily counts are lower than they had been in early May, RR is trending upward in several Central Ohio counties, due in part to having had few cases previously and there being relatively fewer cases elsewhere in Ohio during the current screening period.

Examples of the different user interface choices and corresponding output for Franklin County:

Further Detail on User Input Choices (left panel):

*Aggregate Variable:*

For analytical comparisons, the user may choose state, zone, or region as the baseline variable.

*Aggregate level:*

For analytical comparisons, the user may choose which zone or region to compare within. If one choses the default, ‘State’ as the Aggregate Variable, then ‘OH’ is chosen automatically as ‘Aggregate level.’ ‘Zone’ yields three potential user input options, and ‘Region’ yields eight, according to the nested geographical attributes of counties for state planning purposes.

*Date:*

Users may slide horizontally to choose the date of daily new cases to be mapped in the adjacent upper right panel. Users may also use the play triangle to scroll through a moving mapping of daily new cases, however the zoom will reset each date, so this is optimized for view the state.

*Counties:*

Users may type in which counties are to be mapped from those counties contained in the chosen aggregate level. Ohio has 88 counties. Zone 1 contains Regions 1, 2, and 5 (36 counties) in northern Ohio. Zone 2 contains Regions 4, 7, and 8 (36 counties) in central/eastern Ohio. Zone 3 contains Regions 3, and 6 (16 counties) in southwestern Ohio. In one wishes to map several counties from separate Regions/Zones, presently one must choose ‘State’ and ‘OH’ in the top two options.

*Log Scale*

UNCHECKED by default. This preforms log transformation to the output (y-axis) plotted in the lower right pane. This is base 10 for daily new counts and base 2 for Relative Risk. This can be helpful in number of ways given the range in daily new cases and corresponding relative risks to be plotted.

*Confidence band?*

CHECKED by default, where the shaded ribbon represents the confidence band. When checked, a shaded area indicating a 95% credible interval for the fitted line is displayed. More detail on this aspect is given in the Statistical Methods section below.

*Relative Risk (RR)?*

UNCHECKED by default. In the default, the user observes a classical epidemic curve of daily new cases in the y-axis against date in the x-axis. The user-input CHECK yields a crude relative risk plot (points and fitted curve) over time, where those in the chosen county or counties are the comparable to the those in the chosen state, zone, or region chosen at the aggregate level. With the box CHECKED, the plotted relative risk does take into account population of the given geographical area chosen (unlike the daily new counts, which is not population scaled). More detail on this aspect is given in the Statistical Methods section below. Further introductory resources on Relative Risk calculations are available from the CDC [here](https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section5.html).

*Focus area excluded from RR?*

CHECKED by default, where RR is relative risk. When UNCHECKED, the total cases/population within the aggregate level comparison group is used in calculating the expected cases; i.e. the comparison also contains cases from the chosen county or counties which are being compared. When CHECKED, these are left out such that the plotted crude relative risk represents the risk within chosen counties compared to the risk in the chosen aggregate level without the chosen county. If one chooses *Aggregate Variable*= ‘Region’, then *Aggregate Level* = ‘4’, then adds more than one of the counties in Region 4 (see above map), one can more fully appreciate this feature of RR as constructed here. When CHECKED, as more counties are selected RR becomes increasingly large, approaching infinity. This is because eventually the chosen group becomes all of the cases observed, and none occur in of the comparison portion of the Region used to generate cases expected. When UNCHECKED, as counties are added generally the RR trends towards 1; if all counties are selected, the crude RR is exactly 1 for all observed timepoints because both the chosen group of counties and comparison group of counties are the same. More detail on this aspect is given in the Statistical Methods section below.

# Statistical methods

## Relative risk model

The primary analysis model constructs longitudinal estimates of relative risks in a *focus area* versus the rest of an *aggregation area*. For example, if the focus area is Franklin County and the aggregation level is Region 4, we compute the risk in Franklin County relative to Region 4 less Franklin County. The *crude relative risk* at each date is computed as follows:

* The *expected count* is the sum of the counts in all counties in the aggregation area less any counties in the focus area, multiplied by the ratio of the sum of populations in the focus area over the sum of populations in all counties in the aggregation area less any counties in the focus area;
* If the expected count computed above is zero, it is replaced by an expected count of 1 / 100;
* The *observed count* is the sum of all the counts in all the counties in the focus area;
* The crude relative risk is the observed count divided by the expected count.

If a plot of relative risks is requested, the crude relative risks are plotted as points.

The model fit is a generalized additive model (GAM) with observed count as the outcome with a negative binomial likelihood and a penalized thin plate regression spline term for the date. The log expected count is used as the offset. The negative binomial likelihood is similar to the Poisson likelihood but includes a dispersion parameter allowing for overdispersion relative to the Poisson distribution. The model is fit using the gam function of the mgcv package in R with the default settings:

gam(Count ~ 1 + offset(log(Expected)) + s(Date),

family = nb(), data = data)

If an error is thrown while fitting the model (generally due to too few non-zero observations to support the regression spline, the regression spline term is replaced by a linear term. Since the log link function is used for the negative binomial likelihood, a linear trend in time on the linear predictor scale corresponds to an exponential trend (either growth or decay) on the outcome scale.

If a plot of relative risks is requested, the predicted relative risk curve over time is plotted, along with a ribbon representing the pointwise 95% posterior credible band. A posterior credible band is the Bayesian counterpart to a confidence band.

Note: a lag in case reporting has no average effect on the relative risk model as long as the lag is uniform across geographic areas.

## Daily count model

The daily count model is similar to the relative risk model, except no offset is included. The daily count model is fit to the focus area only. In particular, no spatial smoothing is employed.

If a plot of daily counts is requested, the daily counts are plotted as points, the predicted daily count curve over time as a curve, and the pointwise 95% posterior credible band as a ribbon.

Note: a lag in case reporting generally induces an apparent downward trend near the end of the observation period.

## Potential Alerts

All focus areas of a certain level of the geographical hierarchy within a chosen aggregation area may be screened for a relative risk greater than 1, an increasing relative risk over time, or both. For example, all counties in Ohio may be screened. Relative risks and rates of change of relative risks are estimated at the last available date. The rate of change of relative risk is defined as the ratio of the relative risk at the last available date to the relative risk one day prior. These relative risks are the fitted / smoothed relative risks, not the crude relative risks.

For both the relative risk and the rate of change of relative risk, the following quantities are reported:

1. Estimate: the posterior mean;
2. Lower: approximate 2.5% posterior quantile;
3. Upper: approximate 97.5% posterior quantile;
4. Prob(measure > 1): approximate posterior probability that the measure (relative risk or rate of change) is greater than 1.

Note that Lower and Upper form a *two-sided* 95% posterior credible interval, so that Lower will be less than 1 when Prob(measure > 1) is equal to 0.95, and equal to 1 when Prob(measure > 1) is equal to 0.975.

The quantities Lower and Upper are computed by adding to log(Estimate) the 2.5% and 97.5% standard normal quantiles, respectively, times the posterior standard deviation of log(measure) reported by gam, then exponentiating. The posterior standard deviation is the Bayesian counterpart to the standard error. Similarly, Prob(measure > 1) is approximated by the standard normal cumulative distribution function of log(measure) divided by its posterior standard deviation. Posterior standard deviations for the log rate of change of relative risk are computed by pre- and post-multiplying the posterior covariance matrix by the contrast subtracting the second-to-last date from the last date, then taking the square root.

Examples of viewing potential alert procedures include, but are not limited to:

1. Flagging all counties within the state for which there is greater than 95% posterior probability that the rate of change of the relative risk (versus the rest of the state) is positive;
2. Flagging all census tracks within a county for which there is greater than 95% posterior probability that the relative risk (versus the rest of the county) is positive.

This alerting tool is not meant to indicate that there is a problem at flagged locations. Rather, it is meant to draw attention to flagged areas for closer inspection by experts. Areas may be flagged due to real problems (e.g., increased community transmission), or due to record-keeping (e.g., clearing a backlog of lab tests) or otherwise known reasons (e.g., comprehensive testing of a group care facility).