# Disease Outbreak Radar: A Tool for Public Health Users

Cloris Feng (clorisf@student.ubc.ca), Tae Yoon Lee (taeyoon.harry.lee@gmail.com), Derek Tam (derek.tam.94@gmail.com)

## 1. INTRODUCTION

Disease monitoring is an important function of the public health system globally. With the ongoing COVID-19 pandemic, public interest in disease monitoring has risen dramatically in the past months, and various dashboards visualizing disease count data have been made available by government agencies, public health organizations, universities, and hobbyists alike. The dashboards have been developed for multiple purposes: didactic purposes for general public and specific purposes to help public health users effectively and efficiently detect trends and make informed decisions.

In particular, a critical task for public health users is to detect disease outbreaks, where a disease outbreak is defined as "the occurrence of disease cases in excess normal expectancy" [1]. COVID-19 is one of more than 60 notifiable diseases or disease categories being monitored in British Columbia, and these diseases exhibit a variety of trends, such as seasonality in meningococcal disease and a gradual decline in acute hepatitis A. Consequently, it creates a great burden on public health users, who need to consider these complex disease characteristics and trends when deciding whether a reported disease count is higher than expected. Inevitably, the detection process is both fallible and costly. To alleviate the burden on public health users, an automated method has been developed in collaboration with the British Columbia Centre for Disease Control (BC CDC) [2].

The goal of this project is to develop a visualization dashboard specifically designed for public health users to help them efficiently process the output of the automated method for the main task of detecting disease outbreaks. The rest of the report is organized as follows. In Section 2, we will review several dashboards for detecting outbreaks. In Section 3, we will describe the data and task abstraction for this project. In a subsequent section, we will analyze existing COVID-19 dashboards to identify useful aspects and idioms for designing our dashboard.In Section 5, we will integrate the recommended guidelines and idioms to design and implement a dashboard. We will present our results in Section 6 and discuss limitations and future work in Section 7.

## 2. RELATED WORK

### 2.1 Public Health Intelligence for Disease Outbreaks (PHIDO) Dashboard

The PHIDO dashboard is an application software developed at the BC CDC to help the public health users detect disease outbreaks using the BC CDC's in-house algorithm (referred to as the PHIDO algorithm) for internal use. It allows the users to 1) upload their own dataset in a tabular

format with two columns, the number of disease cases and the date at which the cases are collected, 2) specify the setting of the algorithm (e.g., maximum time, convergence criterion) and the categorization of the outbreak or alert level based on a p-value that measures the probability of observing a case based on their algorithm, and 3) visualize and inspect the output of the algorithm.

For visualization, the PHIDO dashboard uses a 2D line plot to show the observed disease count cases across time (Figure 1). A circle mark is used to encode the cases, and the cases are connected by a dotted line to encode the sequential nature of the data. Color is used to encode the alert level of a disease count into three categories: high (red), medium (yellow), and low (green). The categories of the alert levels are predetermined by the users in the setting. A solid, blue curve is the disease count curve estimated by the algorithm, and it is superimposed on the same plot for the users to compare and investigate the validity of the algorithm and its output. The PHIDO dashboard also provides a filtering option to choose a specific time window and zoom in for closer inspection.

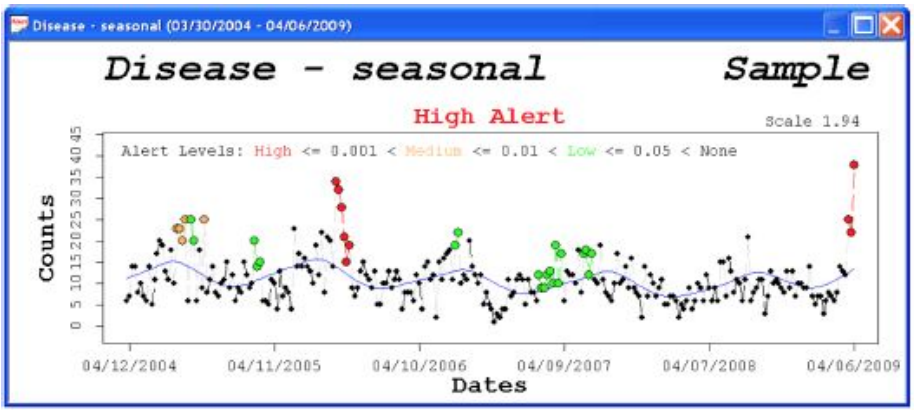


**Figure 1.** Disease outbreak detection visualization in the PHIDO dashboard (Figure 3.1 in [2])

### 2.2 British Columbia Asthma Monitoring System (BCAMS)

BCAMS collects forest fire smoke exposure (e.g., $PM_{2.5}$) and asthma cases for public health studies on the association between the smoke exposure and abnormal rises in asthma cases. McLean et al. used the PHIDO algorithm to define asthma excursions and studied the association between the excursions and fire smoke exposure during the 2014 Forest Fire Season [3]. The authors used a bar chart to encode asthma counts across time. Similar to the PHIDO dashboard, the color was used to encode the type of asthma excursions: rare (red), unusual (yellow), normal (green). The asthma count curve estimated by the PHIDO algorithm was superimposed on the same plot as a black, solid curve. For the final result, the authors aggregated the asthma cases as the proportion by the excursion type and used a stacked bar chart to show the percentage of asthma excursions across the four levels of the PM2.5 concentrations, with the color encoding the type of excursions.

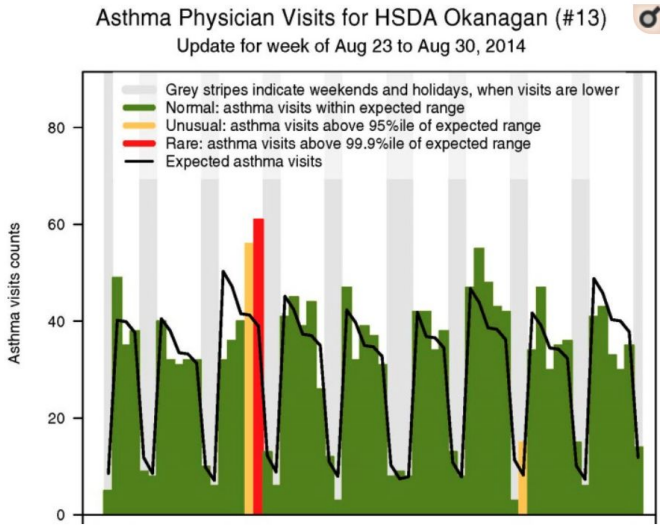


**Figure 2.** Asthma cases and excursions during the 2014 Fire Forest Season (July 1, 2014 - August 30, 2014) in Okanagan, British Columbia (Figure 1 in [3]).

### 2.3 Somalia Polio Room Dashboard

Kamadjeu et al. developed the Somalia Polio Room Dashboard to track polio outbreaks in Somalia, Kenya, and Ethiopia during 2013-2014 [4]. The design aims to address the needs of timely information on cases and performance indicators of the polio outbreak and to provide decision makers with a graphical display of the needed information. The authors utilized composite graphs, maps, and tables to show the trend and geographical information of polio cases, the key factors of polio cases, and essential performance indicators for Acute Flaccid Paralysis surveillance (Figure 3). The authors indicate that the Somalia Polio Room Dashboard is able to provide users with efficient information display and integration.

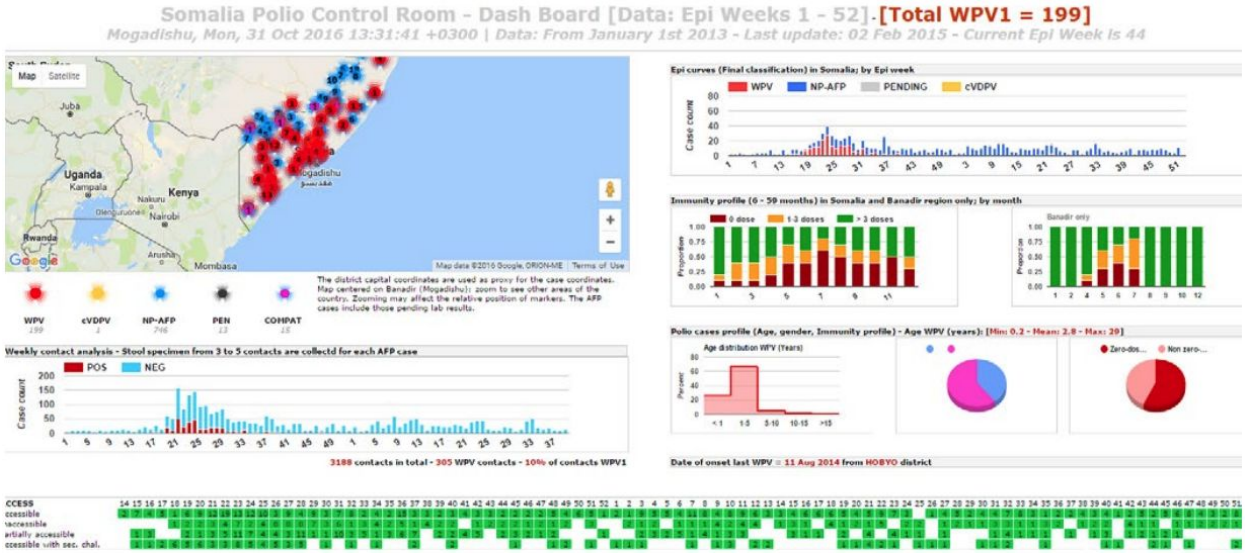

**Figure 3.** Somalia Polio Control Room Dashboard (Figure 4 in [4])

## 3. DATA AND TASKS

### 3.1 Data

Disease count data consist of spatio-temporal aggregations, providing the number of disease cases over a region during successive, regular time intervals. For this project, a publicly available disease count database of the United States (US) from Project Tycho was used [5]. After preprocessing the database, a clean subset was obtained with no missing values. There were 52 states with 15 diseases, each of which being collected on a weekly basis for each state from 2009-01-04 to 2014-07-27. This translated to 290 data points for each disease in each state. Description of the finalized database is provided in Table 1.

| **Attribute name** | **Attribute type** | **levels/range** | **Description** |
|---|---|---|---|
| Disease type | Categorical | 15 | Disease type |
| Week range | Sequential/ ordered | 2009-01-04 - 2014-07-27 | Start and end dates of a week in which the number of cases is collected (290 observations) |
| US state | Categorical | 52 | US states |
| Number of cases | Quantitative | 0 - 3140 (integers) | Number of cases of a disease |

**Table 1**. Description of the finalized disease count database.

We now describe the derived input and output attributes for the automated method. The automated method took three inputs: the number of cases, the week number, and two trigonometric bases. The week number is the number of weeks starting from 2009-01-01 (so that it is equal to 1 on 2009-01-07), and it was used to account for both the long-term and short term time trends in the disease count. The trigonometric bases were used to incorporate the seasonal pattern of the disease count. Note that the method was run on each disease for each state.

The automethod method essentially provides a robust curve of the disease count by down-weighting the counts that are abnormally high. The direct output of the method is an 'robust' estimated number of cases at each of the week numbers. Based on the method, for each week, the probability of obtaining the number of cases at least as extreme as the observed number of cases was computed; this is referred to as the (individual-level) p-value. Then an alert level for a disease count was derived from the p-value. The levels need to be defined by the user. For instance, the BC-CDC uses three levels: low (0.01<p-value <= 0.05), medium (0.001 <

p-value <= 0.01), and high (p-value < 0.001). For a measure of a disease outbreak, we computed the outbreak p-value based on the last ***m*** number of reported cases, where ***m*** needs to be specified by the user. For instance, the BC CDC uses m=3. Then similarly, an outbreak level was derived from the outbreak p-value according to the categorization defined by the user. Table 2 summarizes the derived attributes.

| **Attribute name** | **Attribute type** | **Levels/range** | **Description** |
|---|---|---|---|
| **Input attributes** | | | |
| Week number | Quantitative | 1-290 (integers) | Week number starting from 2009-01-07 |
| Trigonometric bases | Quantitative (cyclic) | cos(2 pi 7*t/365.25), sin(2 pi 7*t/365.25) | Trigonometric bases based on the week number |
| **Output attributes** | | | |
| Estimated number of cases | Quantitative | Non-negative real valued | The number of cases estimated by the automated method |
| p-value | Quantitative | 0-1 (real-valued) | Probability of obtaining the number of cases at least as extreme as the observed number of cases based on the automated method |
| Alert level | Categorial | User-defined | Level of alert specified by public health users based on the p-value. For example, an alert level is low, medium, or high if the outbreak score is 0.01-0.05, 0.001 - 0.01, or less than 0.001, respectively. |
| Outbreak p-value | Quantitative | 0-1 (real-valued) | Probability of obtaining the sum of the three most recent number of cases at least as extreme as the sum of the three most observed number of cases based on the automated method. |
| Outbreak level | Categorical | User-defined | Level of outbreak specified by public health users based on the outbreak p-value. |

**Table 2.** List of derived attributes for the automated method.

### 3.2 Task Description

There are three main tasks that public health users are most interested in:

- **T1.** As the number of diseases to monitor across the regions is over 700, one important task for public health users is to efficiently **search** which disease is at high risk of an outbreak in each region. This task is performed each time a new aggregated disease count is reported (for instance, it is weekly for the BC-CDC), contributing to a significant burden on the public health users.
- **T2.** If a disease with a high outbreak level is found, then they need to **analyze** the trend in the reported cases of the disease as well as the output of the automated method, so as to confirm the validity of the outbreak level. For example, the BC CDC is most concerned with the last three weekly disease counts, from which the outbreak level is computed.
- **T3.** Moreover, the users **compare** the outbreak level of a disease across the regions and investigate whether there is any spatial pattern.

## 4. Public Dashboards for Virus Spread

As the COVID-19 pandemic develops around the world, dashboards for tracking information on COVID cases and case development trends are available online for public use. In this project, we plan to implement a dashboard to track information on disease outbreak levels and the trend of disease case counts for a US dataset. In this analysis, we reviewed a few COVID dashboards and extracted features and idioms from them, which were useful and inspiring in helping us come up with ideas on our dashboard design. Even though there are differences in tasks between COVID dashboards and ours, we still learned some useful lessons from the existing COVID dashboard implementations and planned to apply them to our design. More specifically, the lessons that we learned from this analysis will be useful in helping us address T2 and T3 in the task abstraction.

***Overview Design on Reported Case Counts of A Specific Disease***

In some existing COVID-19 dashboards [1], [6]–[8], aggregated confirmed COVID-19 cases are presented in the form of a choropleth, such as the one shown in Figure 4, where the numbers of reported cases across the states are scaled in the different degrees of color change. The redder a specific region on the map is, the more confirmed cases are reported in that region. This form of representation is beneficial for users to identify the regions with most reported cases by judging the hues shown on the map. However, the data in Figure 4 is collected at the county level, which would be visually heavy for users to spot the states with most accumulated cases. In this project, we could use a choropleth to show the aggregated cases for a specific disease across all states so that users could quickly identify the regions with most reported cases, which is relevant to the need of T3 of this project. Also, since our data was collected at the state level, which could

lessen visual burden for users while showing the overview of a specific disease across all regions, a choropleth for reported confirmed cases could be used to provide users with information on confirmed cases of a specific disease across all regions. However, one thing worth noting here is that the severity of how COVID spreads over the population within a specific region isn't addressed with the total confirmed cases. Therefore, some COVID dashboards [1], [8] provide users with choropleths of total confirmed cases by population as well, such as the one shown in Figure 5. We could see from Figures 4 and 5 that there is a shift of regions highlighted in red, which validates the idea that only showing a choropleth of total confirmed cases is not enough to illustrate the severity of the spread of a disease on the population in an area. Figure 5 provides users with information on regions where the population is affected most by the disease. This could be included as an option for the dashboard users in the project implementation.

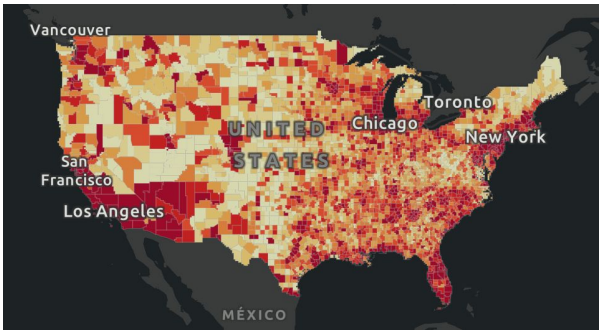


**Figure 4**. Aggregated confirmed COVID cases in the United States are illustrated in a choropleth at the Johns Hopkins Coronavirus Resource Center website [8].

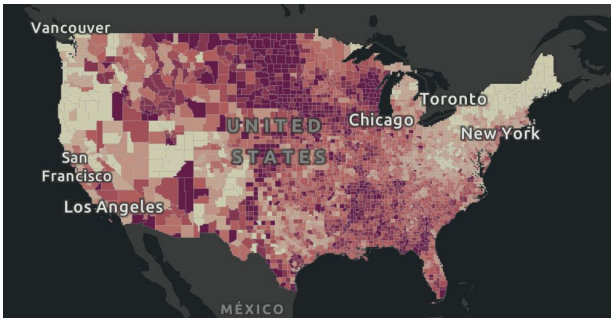


**Figure 5**. Aggregated confirmed COVID cases by population in the United States are illustrated in a choropleth at the Johns Hopkins Coronavirus Resource Center website [8].

***Overview Design on Disease Outbreak Levels of A Specific Disease***

Besides choropleths, some COVID dashboards use other idioms to show total cases reported in the US. For instance, the bubble map shown in Figure 6. The size of bubbles on the map represents the amount of total COVID cases in a specific state. The larger the bubble is, the more

reported cases are within the specific region. Also, when users hover over a specific bubble in a region, a box will appear and show the total, active, recovered, and fatal case counts in that region and each case count category is represented with a colored circle legend for better visualibility. The usage of bubble size in this representation is not of high distinguishability, when the amounts of case counts do not have large differences, leading to less differences in bubble size and more difficulties in judgement by eyes. However, in this project, a similar bubble map design could be used to show an overview of disease outbreak levels of a specific disease across all regions. In our project, we would have low/medium/high outbreak levels to classify the risk of disease outbreak in a specific region. In this case, we could use different colored bubbles, such as green/yellow/red bubbles to represent low/medium/high outbreak levels of a specific disease. Instead of having orange bubbles shown on map as illustrated in Figure 6, we will show the outbreak level of a specific disease in a region by showing its corresponding colored bubble within the specific state. In this way, the user could quickly identify states at a high risk of disease outbreak by finding states with red bubbles shown on the map and this representation could meet the need of T3 of this project.

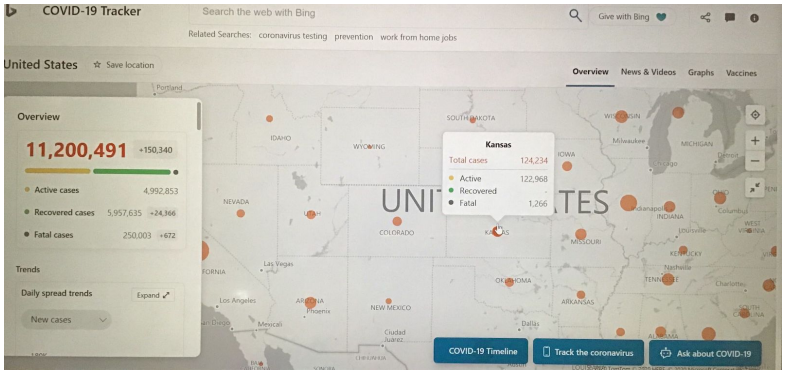


**Figure 6**. An overview of total cases in the United States shown in a map at the Bing COVID-19 tracker dashboard [6].

***Trend of Reported Case Counts***

There are different idioms in COVID dashboards that show the trend of reported case counts. In this section, we will analyze and extract and combine features from these idioms and find the design that is more suitable for T2. In Figure 7, a bar chart is used to show the trend of confirmed cases reported weekly in the United States. Also, if users hover over a bar, they will have more information on the weekly change of case counts compared with the reported cases from the previous week. In contrast, Figure 8 shows a line over bar chart, which illustrates new daily cases reported in a state with a bar chart and the 7-day moving average of the new COVID cases with a line on top. More specifically, the green segment of the line indicates a day-over-day decrease in new cases while the red segment of the line means a day-over-day increase in new cases. If there is no change, the segment of the line will be grey.

In our project design, we could combine features from these two idioms. We could use a line over bar chart to show the trend of weekly reported case counts of a specific disease in a specific region, where the bars will represent weekly reported cases and the line will be the curve showing the trend of reported cases. Specifically, the curve will be color-coded to indicate the trend of increase, no change, or decrease of weekly reported cases similar to what is shown in Figure 8. The green segment of the line indicates a week-over-week decrease in reported cases while the red segment of the line means a week-over-week increase in reported cases. If there is no change in case numbers, the segment of the line will be grey. When users hover over a specific bar, they will have an information box, similar to what we've seen in Figure 7, which shows reported case counts in a specific week and the percentage of change compared to reported case amounts from the previous week. This idiom implementation is able to address the task stated in T2, while providing more details, such as percentage of change in case numbers, to users if they are interested.

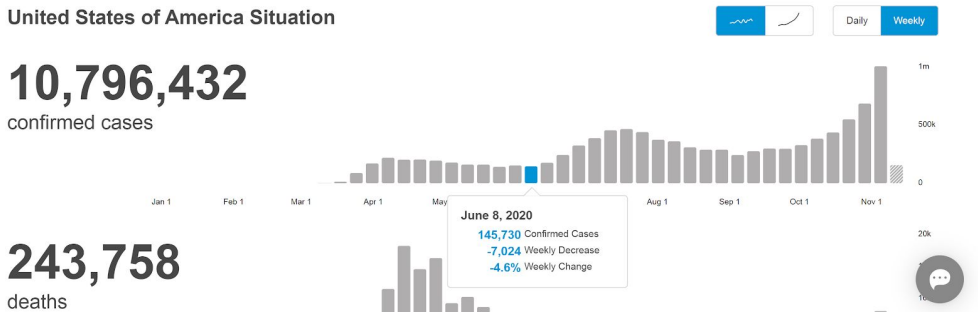


**Figure 7**. Trend of total confirmed cases reported weekly in the US shown at the WHO website [1].


**Figure 8**. Trend of new daily COVID cases reported in the state of Illinois and the trend of 7-day moving average [7].

***Analysis Summary***

Here is a summary of all key points mentioned in previous analysis sections, which would address the need of T2 and T3:

1) For **T2**:
    a) Use a line over bar chart to show the trend of weekly reported case counts of a specific disease in a specific state, where the bars will represent weekly reported cases and the line will demonstrate the trend of reported cases.
    b) The line will be color-coded to indicate the trend of increase or decrease change of weekly case counts. The green segment of the line indicates a week-over-week decrease in reported cases while the red segment of the line means a week-over-week increase in reported cases. If there is no change in weekly reported cases, the segment of the line will be grey.
    c) When users hover over a specific bar, they will have an information box, which shows reported case counts in the week and the percentage of change compared to cases from the previous week.
2) For **T3**:
    a) Use a choropleth to show the aggregated cases for a specific disease across all states so that users could quickly identify the regions with most reported cases.
    b) Add a choropleth of total confirmed cases by population to show the severity of the spread of a disease on the population in an area.
    c) Use a bubble map to show the outbreak level of a specific disease in a region by showing its corresponding colored bubble within the specific state.
    d) On the map, green/yellow/red bubbles represent low/medium/high outbreak levels of a specific disease.

## 5. PROPOSED SOLUTION

The final goal of this project is to produce a dashboard for internal use at public healthcare systems, such as the BC CDC, to determine outbreak status for diseases under monitoring. Based on the compiled and summarised recommendations from our review combined with our observations on publicly available dashboards, we will adjust design decisions for the dashboard proposed below.

### 5.1 Overview

Two prototypes have been compiled through collaborative discussion and assessment of the core user task abstractions. It is obvious that a combination of visualizations will be required to both search for and focus on points of interest, which warrants the use of a dashboard-style assemblage of views that together are effective in completing the defined user tasks. To this end,

two prototypes have been created that focus on different aspects of the tasks, which are discussed below.

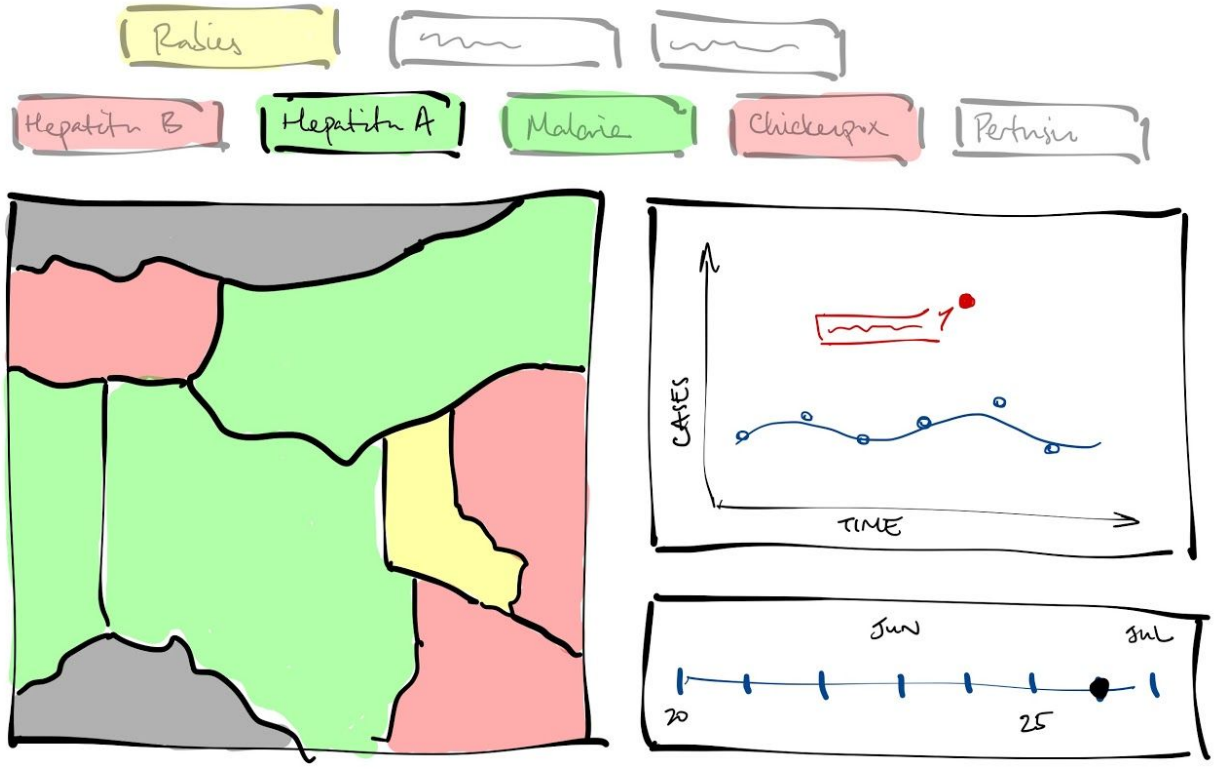


**Figure 9.** Sketch of proposed dashboard prototype #1 prior to aggregation and integration of recommendations.

- **Choropleth map (left):** Will be used as a high level overview for the current outbreak risk of a single disease. This would double as a selection tool for displaying the data on the scatterplot. There are known shortcomings with choropleth maps, but this is currently the best compromise for a broad visual overview combined with a navigational interface.
- **Scatterplot with fitted curve (top right):** Will serve as the visualization of the outbreak model as a curve with points representing previous case numbers in the past time window. Details for a specific datapoint can be investigated with a mouse-over to determine the exact number of cases and the calculated likelihood of that case being an outbreak with the associated statistics.
- **Timeline (bottom right):** Will serve as a selector for the time period that the user would like to investigate, and could be "scrubbed" to move back in time which will be reflected in the choropleth map and time series scatter plot.

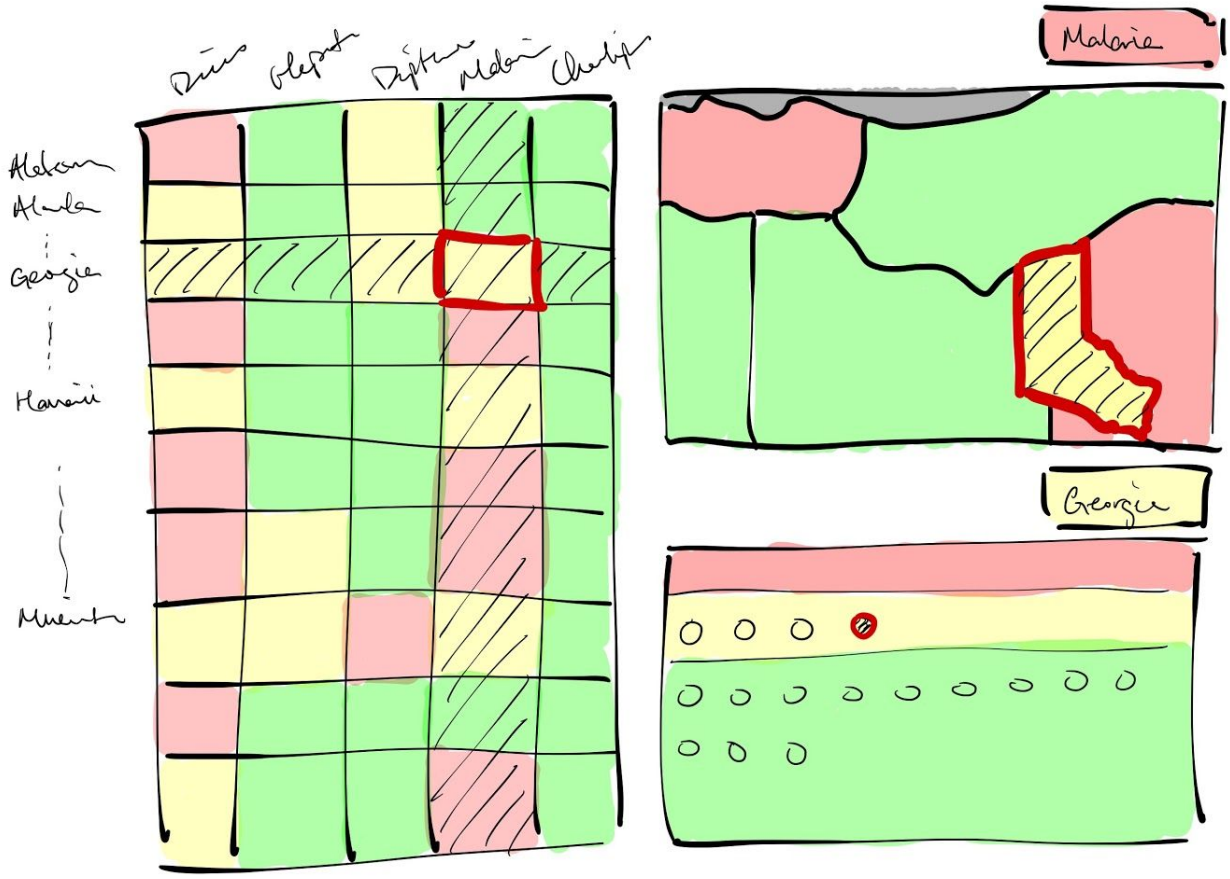


**Figure 10.** Sketch of proposed dashboard prototype #2 prior to integration of recommendations.

- **Heatmap (left):** Serves as a high level overview of outbreak statuses for all diseases across all monitored regions. Depending on the cardinality of the data, this heatmap may become difficult to interpret as the information density increases, but different magnification techniques can be applied to augment usability. The heatmap would double as a selecting tool for highlighting a disease and region of interest simultaneously, which will be reflected in the choropleth and grouped tabular marks on the right.
- **Choropleth map (top right)**: As in prototype #1, the choropleth provides visual intuition for the spatial organisation of regions of interest for a specific disease.
- **Grouped tabular summary (bottom right):** Acts as a visual summary for the outbreak status of all monitored diseases in a region of interest.

### 5.2 Detailed Time-series View

When a user selects a specific disease in the heatmap or the grouped tabular summary, a detailed time-series will be presented on the right in stacking faceted panes as more diseases are investigated.

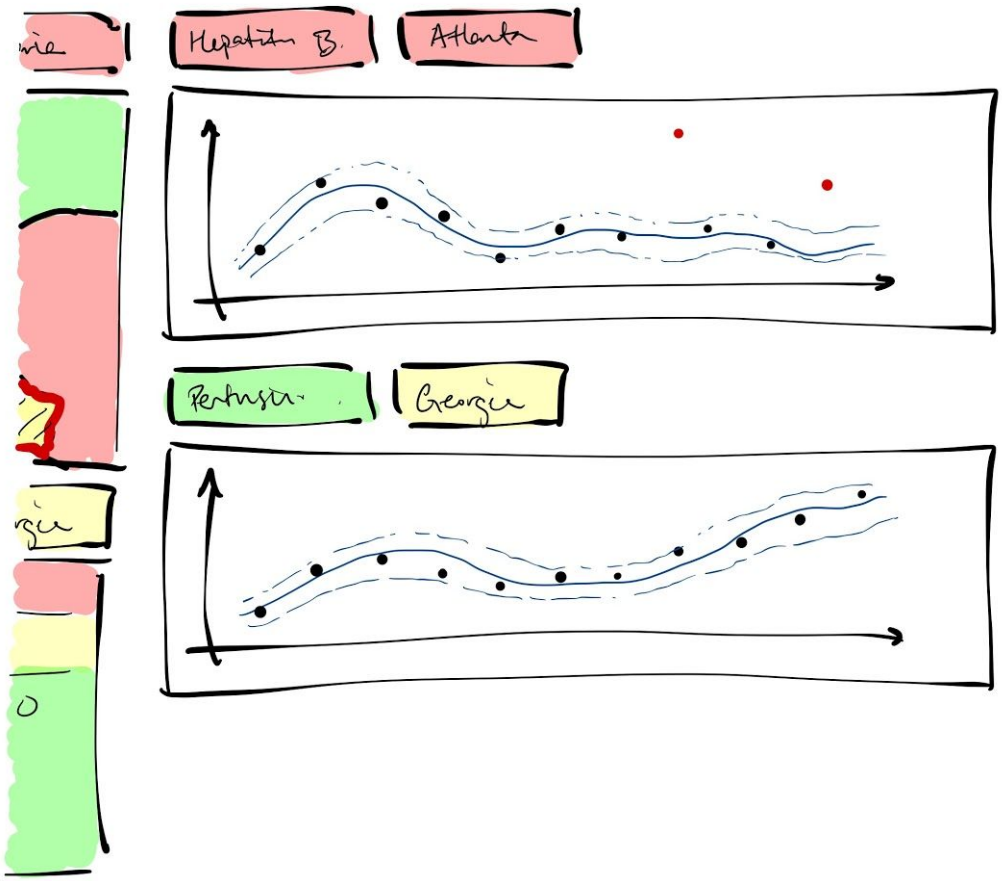


**Figure 11.** Sketch of proposed detailed view for specific diseases of interest.

- **Time series scatter plots:** Presents the specific periodic case data for a region and disease over time. A trendline is included that represents the ***robust fit line*** from the statistical model. Observations that are found to be high or medium outbreak risk are colored consistently with the overview. Details can be found for any given data point on mouse-over.

### 5.3 Implementation

The final dashboard will be deployed through an interactive R ShinyApp, as most public health users are familiar with that platform. The interactive aspects of the dashboard will be animated using D3.js and executed within R using the R2D3 package.

## 6. RESULTS

### 6.1 Proposed use case

## References


[1] "WHO | Disease outbreaks," *WHO*. https://www.who.int/environmental_health_emergencies/disease_outbreaks/en/ (accessed Oct. 22, 2020).

[2] T. Y. (Harry) Lee, "Robust methods for generalized partial linear partial additive models with an application to detection of disease outbreaks," University of British Columbia, 2019.

[3] K. E. McLean, J. Yao, and S. B. Henderson, "An Evaluation of the British Columbia Asthma Monitoring System (BCAMS) and PM2.5 Exposure Metrics during the 2014 Forest Fire Season," *Int J Environ Res Public Health*, vol. 12, no. 6, pp. 6710–6724, Jun. 2015, doi: 10.3390/ijerph120606710.

[4] R. Kamadjeu and C. Gathenji, "Designing and implementing an electronic dashboard for disease outbreaks response - Case study of the 2013-2014 Somalia Polio outbreak response dashboard," *Pan Afr Med J*, vol. 27, no. Suppl 3, p. 22, 2017, doi: 10.11604/pamj.supp.2017.27.3.11062.

[5] W. G. van Panhuis *et al.*, "Contagious diseases in the United States from 1888 to the present," *N Engl J Med*, vol. 369, no. 22, pp. 2152–2158, Nov. 2013, doi: 10.1056/NEJMms1215400.

[6] "Microsoft Bing COVID-19 Tracker." http://bing.com/covid/local/maine_unitedstates?dynamicSharing=true&shtp=Facebook&shwth=900&shh=800&shtk=Y29yb25hdmlydXMgdHJhY2tlciB1cGRhdGVz&shdk=dGVzdA%3D%3D&shth=OSH.Mmq%2BwuM5WWl/TcdViNGxBA&redirect_uri=http%3A//veeraux%3A81/covid/local/unitedstates%3FdynamicSharing%3Dtrue&ref=Coronavirus&al (accessed Oct. 22, 2020).

[7] "Track Testing Trends," *Johns Hopkins Coronavirus Resource Center*. https://coronavirus.jhu.edu/testing/tracker (accessed Nov. 17, 2020).

[8] "COVID-19 Map," *Johns Hopkins Coronavirus Resource Center*. https://coronavirus.jhu.edu/map.html (accessed Oct. 22, 2020).