BC Chronic Disease Capstone Report

Capstone Partner: BC Office of the Provincial Health Officer, Mentor: Daniel Chen

Team: Jennifer Hoang, Jessie Wong, Mahmoodur Rahman, Irene Yan

2022-06-22

Contents

1 Executive Summary **2**

2 Introduction **3**
2.1 Temporal Modelling . . . 3
2.2 Dashboard . . . 3
2.3 Join-point Regression . . . 3

3 Data Science Techniques **3**
3.1 Temporal Modelling . . . 4
3.2 Dashboard . . . 4
3.3 Joinpoint Regression . . . 5

4 Data Product and Results **5**
4.1 Temporal Modelling . . . 5
4.2 Dashboard . . . 5

5 Conclusion and Recommendations **10**

References **11**

1 Executive Summary

The BC Chronic Disease Registry (BCCDR) is a data product that captures information about the rates of new and persistent cases of 25 different chronic diseases across the Province of British Columbia. Crude and age-standardized rates of disease are recorded for different health boundary types, including HAs (Health Authorities) and CHSAs (Community Health Service Areas), as well as for demographic variables such as sex. In this project we aim to develop an analysis pipeline to describe the temporal trends in the data, and then build an interactive dashboard that will allow users of all technical expertise to explore and visualize spatial and temporal information of the disease rates. This report will outline the approach we took to tackle this problem and a description of our final data product.

2 Introduction

Millions of people in BC live with a chronic disease, so it's important to understand and interpret the distribution of diseases throughout the province for a variety of reasons. We may want to know how to best allocate healthcare resources, or to identify if a specific region is experiencing rapid growth of a disease. The dashboard is a tool that will allow healthcare professionals and eventually the general public to access the disease information and answer these questions.

The BC CDR captures 3 different types of rates that we will be incorporating into our dashboard. Incidence Rate is the rate at which new cases occur in a specified population during a specified time period; Lifetime Prevalence is the proportion of individuals who have had the condition for at least part of their lives, and Active Healthcare Contact Prevalence are the cases for which a patient seeks healthcare services for relapsing - remitting conditions. Each of these 3 rates are available as Crude or Age Standardized rates. Age Standardized rates are adjusted to the standard population, and are calculated as if all regions shared the same age structure, that of the 2011 Canadian Census. Crude rates are not adjusted to the standard population, and represent the number of cases in a specific geographic region divided by the population at risk in that region. For each disease rate metric, the data is stratified by year, sex and by region at various tiers. In this project we focus on the least and the most granular health boundaries, which are the 5 Health Authorities (HA) and the 195 Community Health Service Areas (CHSA).

2.1 Temporal Modelling

To help interpret chronic disease trends across the province, we generated models to smoothen disease rate estimates at the most granular health boundary, the CHSA, since estimates from smaller communities can have large fluctuations and confidence intervals. We selected a Bayesian modeling approach and a local polynomial regression (LOESS) model to be fit on data over 20 fiscal years from each combination of disease, CHSA, and age-standardized rate type. These temporal models were integrated into the final dashboard to allow users to easily compare chronic disease rates and trends between different CHSA.

2.2 Dashboard

The dashboard facilitates the exploration and visualization of spatial and temporal trends of 25 different chronic diseases across the Province of British Columbia. The dashboard consists of 4 main tabs, which are Information, By Disease, By Region, and Data. The Information tab contains a description of the dashboard usage, and definitions of rate types, diseases, and data variables. The By Disease tab allows for the comparisons of one disease over several HAs or CHSAs, while the By Region tab allows for the comparisons of several diseases in one particular HA or CHSA. The Data tab retrieves and displays all data specified by the user, and contains a button to download the data if the user wishes to do so.

2.3 Join-point Regression

Segmented or broken-line models are regression models where the relationships between the response and explanatory variable are piece-wise linear, namely represented by two or more straight lines connected at unknown values: these values are usually referred as breakpoints, change-points or even join-points. Here the response variable are different standardized chronic disease rates, and Time in years (2001 to 2020) being the explanatory variable. Broken-line relationships are common in many fields, including epidemiology, occupational medicine, toxicology, and ecology, where sometimes it is of interest to assess threshold value where the effect of the co-variate (Time in years) changes (Ulm 1991; Betts, Forbes, and Diamond 2007). In other words, this model is used here to estimate abrupt changes in rates in particular points in time, rather than smoothing it. The idea is to aid epidemiologists in drawing inferences by estimating these change-points.

3 Data Science Techniques

Several different data science tools and techniques were used throughout the project to accomplish the project deliverable. Some tools were familiar to the team, while others tools were newly learned. The tools and

techniques used in each aspect of the project are described in this section.

3.1 Temporal Modelling

For temporal modelling, we selected the Integrated Nested Laplace Approximation (INLA) approach for Bayesian modelling. INLA is faster compared to Markov Chain Monte Carlo simulation-based methods (Wang (2018)), and since there were over 10,000 combinations of CHSAs, diseases, and rates to be modelled, computational speed was a key consideration. Potential disadvantages of INLA include that small variations between repeated runs can occur as a trade-off for the faster computation gained by using multiple cores (Wang (2018)). However, these small variations are well beyond the significant digits of our data, therefore, we considered this to be an acceptable trade-off.

With INLA, we used a Gamma generalized linear model to smoothen the age-standardized rate, which is a continuous, positive response. The effect of year was modeled using a Random Walk prior to account for temporal autocorrelation. On each CHSA, disease, and rate, a first-order random walk (RW1) and a smoother second-order random walk (RW2) model was fit. The best RW model was selected using the Widely Applicable Information Criterion (WAIC).

A limitation of this Bayesian approach was that rare diseases with low incidence and prevalence rates could not be modelled, because the Gamma distribution does not provide support for values of 0. This most frequently affected small communities and rare diseases, such as juvenile arthritis, multiple sclerosis, and rare forms of stroke. Despite this limitation, our Bayesian smoothing model could be applied to 91% of the data, and more specifically, 84% of incidence rate data, 91% of HSC prevalence rate data, and 98% of life prevalence rate data.

On the remaining 9% of the data containing zeroes, the local polynomial (LOESS) regression baseline model was fit for smoothing. The LOESS regression is a weighted least-squares regression method that considers neighbouring points within a given span. The span was automatically optimized for each set of data using the `fANCOVA::loess.as()` function.

Alternative approaches that were explored to accommodate the data with zero values were to use the Tweedie distribution, which can accommodate positive continuous data with a point mass at zero (Kurz (2017)). However, the implementation of the Tweedie distribution in R-INLA remains experimental and subject to change, and was not further pursued due to reproducibility concerns. Secondly, zero-inflated Gamma and hurdle Gamma models were also investigated to accommodate zero values, but were not supported by R-INLA at this time. Lastly, modification of the zero values into small non-zero values (0.0001) was explored to be used within the Gamma model, however, the RW1 and RW2 models did not demonstrate a suitable level of smoothing after this modification. Further adjustment of the priors and the scaling of the INLA model may be required.

3.2 Dashboard

While the majority of the dashboard was built using R, several other languages were also needed in order to customize and run the dashboard, including JavaScript, HTML, and CSS. R was used to build the Shiny App framework, as well as for all of the data wrangling, processing, and plotting. Javascript was needed to create customized functions that were otherwise impossible in R, such as for the map animation in the By Disease tab. Some Javascript components were also needed to optimize the processing speed of the app. HTML was used for text formatting and other styling options throughout the dashboard, including on the Information pages, in the plot hover labels, and the plot legends. CSS was needed for reusable customized styling, and streamlined the style modification process in the dashboard. Much of the element layouts, colours, and spacing were achieved through the use of CSS.

Nearly 15 different packages were used to create the dashboard, with the most important ones being Shiny, Leaflet, and Plotly. Shiny was used to build the overall app framework, and was a requirement on part of the Capstone Project. Leaflet was used to create the interactive map on the By Disease tab, and was chosen due to its easy integration with Shiny and Plotly. However, the difficulty with using Leaflet is that it does not easily animate within R and Shiny, and thus the custom Javascript functions were needed in this case.

Plotly was used to create the interactive plots in both the By Disease and By Region tab. This plotting library was preferred over other options such as ggplot because of its built-in interactive components, and its ability to create smooth animations via proxies.

3.3 Joinpoint Regression

Often the relationship between the response and explanatory variables is non-linear, where it can be seen that the effect on the response changes abruptly. In case of describing effect of time upon disease rates, epidemiologists are usually interested in the break-point location and the relevant regression parameters. The classical methods used to take into account non-linear effects, such as polynomial regression, regression splines and non-parametric smoothing, are not suitable because the change-points are fixed *a priori* (regression splines) or are not considered at all (smoothing splines and polynomial regression). Moreover, regression parameters obtained in regression splines or polynomial regression approach are not directly interpretable (RJ (1990)). The `R` package, `segmented` (Muggeo et al. (2008)) is capable to account for the unknown breakpoints, and providing fitted values.

We used `segmented.lm` function from the `segmented` package on data consisting of the chronic disease rate over time and CHSA. We performed log-linear regression considering rates as response and time as explanatory variable, then applied the `segmented.lm` function, which gave us fitted estimates, and when plotted shows us the desired breakpoints. A plot from the fitted values of estimated age-standardized incidence rate of Gout in Kimberley CHSA is shown below.

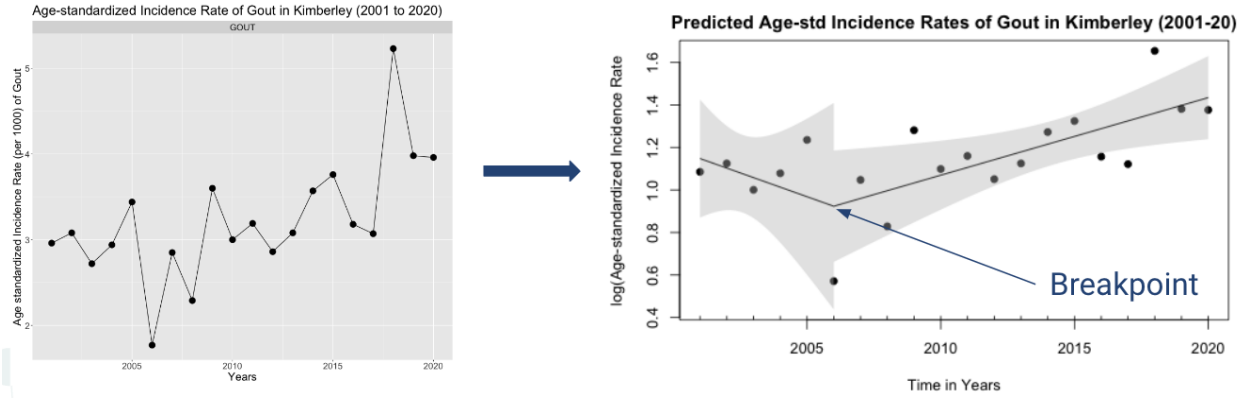

Figure 1: Estimated Age-standardized incidence rate of Gout in Kimberley

Model diagnostics defining how well the data fits the model can be explained by Average Annual Percent Change (AAPC), which is the weighted average of the change in slopes at each join-points between the segmented lines.

4 Data Product and Results

4.1 Temporal Modelling

4.2 Dashboard

The dashboard allows the flexibility for the user to explore and visualize spatial and temporal data of 6 disease rate metrics, for a unique selection of diseases, health boundaries, sex, and year. With the various tabs, the user can make appropriate comparisons as needed to address specific research questions, and can easily visualize temporal trends through animations of multiple data visualizations over time.

The homepage of the dashboard is `About` page of the Information tab, which describes the usage and features of the various tabs on the dashboard. The Information tab consists of three other pages, which include information and definitions of Rate Types, Diseases, and a Data Dictionary. An image of the Information tab is shown below in Figure 2.

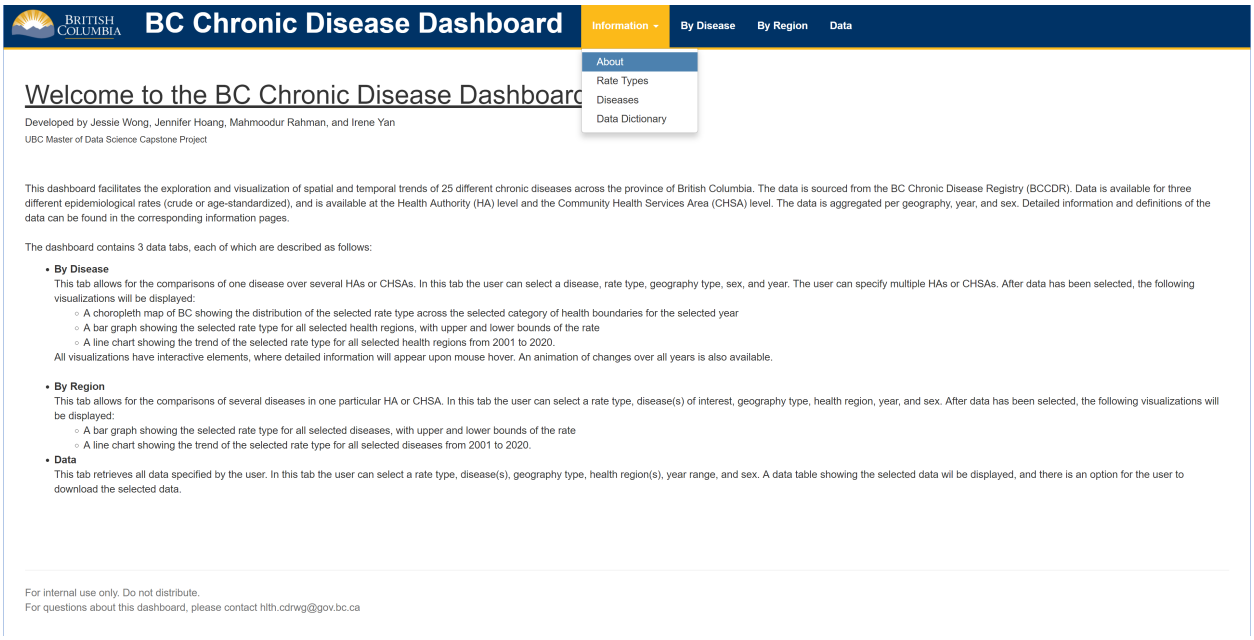

Figure 2: Homepage of the dashboard showing four pages in the ‘Information‘ tab

On the By Disease tab, a user can select a single disease and rate type of interest, multiple health boundaries within a single health boundary type, and a single sex and year. The dashboard subsequently displays the health boundary with the highest maximum and highest average disease rate, the median recorded rate over all health boundaries in the selected health boundary type, and the year of the highest median disease rate. A choropleth map showing the disease rate over the selected health boundary type is also displayed, along with a bar chart showing the rate and confidence intervals of the disease rate for the user selected health boundaries, and a line chart showing the change in the disease rate of the selected health boundaries over time. Additional information is displayed upon hover on any of the plots. The layout of the By Disease tab is shown below in Figure 3.

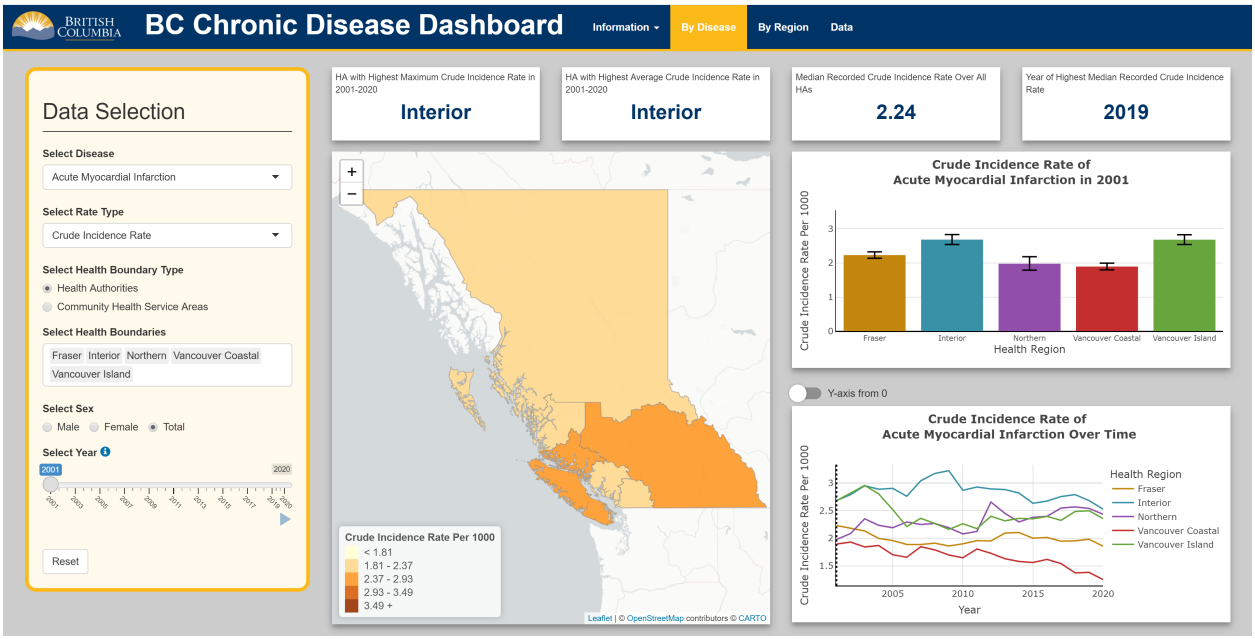

Figure 3: Layout of the 'By Disease' tab of the dashboard, showing summary statistics, map, and two graphs.

On the By Region tab, a user can select a single health boundary within either of the health boundary types, multiple diseases, a single rate type, sex, and year. The dashboard will then display a bar chart of the disease rate and confidence intervals for the selected diseases, and a line chart showing the change in disease rates of the selected diseases over time. The top 4 diseases with the highest rate in the selected health boundary are also displayed. The layout of the By Region tab is shown below in Figure 4.

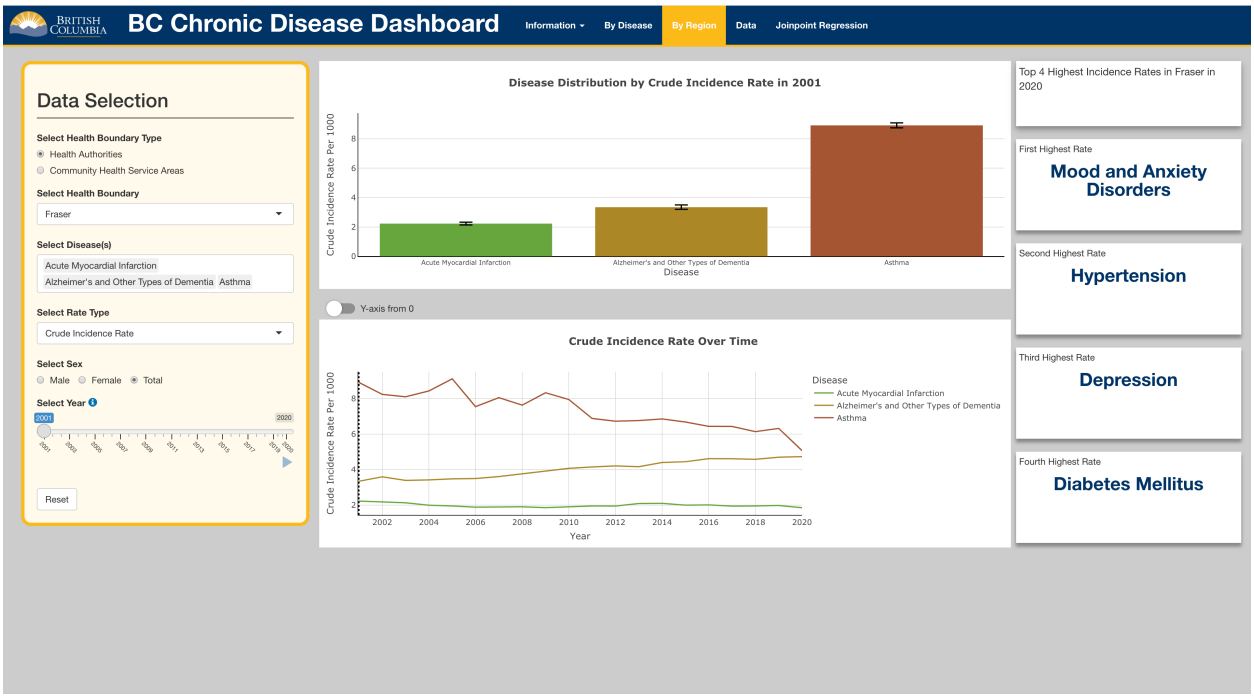

Figure 4: Layout of the 'By Region' tab of the dashboard, showing two graphs, and the top four diseases in a given region.

On both the By Disease and By Region tab, a toggle switch appears above the line chart when modelled data is available, which allows the user to quickly switch between raw data and smoothed time trends. A permanent toggle switch is also present above the line chart that controls whether the y-axis starts from 0 or is adjusted to the range of the data. This customization allows the user to better understand how a disease rate changes over time based on their needs.

Lastly, the Data tab allows a user to select data on multiple diseases and health boundaries, and download the data for further analysis. The layout of the Data tab is shown below in Figure 5.

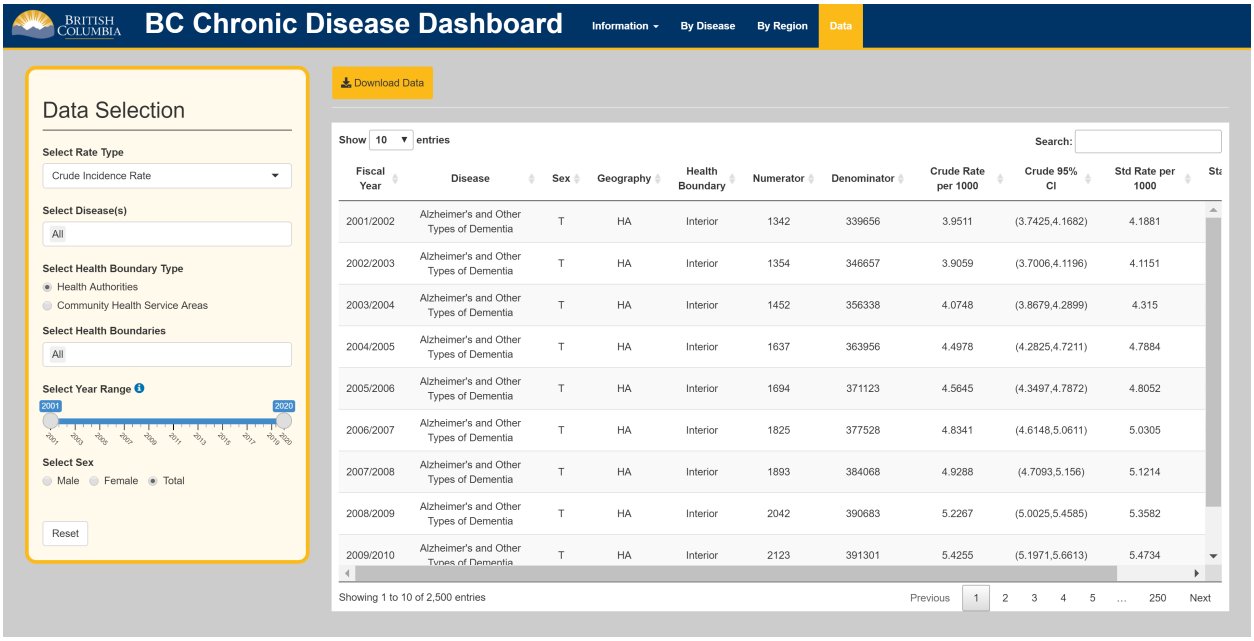

Figure 5: Layout of the 'By Region' tab of the dashboard, showing two graphs, and the top four diseases in a given region.

The creation of this dashboard has the potential to streamline health data extraction and interpretation for epidemiologists and other health professionals, and will increase public awareness of chronic diseases when it is eventually made public. Future improvements that could be made to the dashboard are to increase the processing speed of the dashboard when new data is selected and visualizations are recalculated and reloaded. The dashboard currently runs at an acceptable speed locally, but if the app were to be publicly deployed in the future, there may be some usage concerns regarding loading time. The current start up time of the app is also considerable, because the app needs to read in and wrangle the data in every new session. This issue can be eliminated in the future when the data is made public, since the data can be wrangled once, and the pre-processed data can be safely stored on a server.

The Joinpoint Regression tab of our R Shiny app plots the fitted estimates from the model and also the original data point over a period of 2001 to 2020. There are filters for choosing the Standardized Rate (Response), the Diease, and the CHSA. Toggling these selections, the user can visualize the breakpoints, and also how it reflects the original distribution. The following figure shows the dash board showing the estimated Age-standardized Incidence Rate of Asthma in Kimberley CHSA for the period of 2001 to 2020.

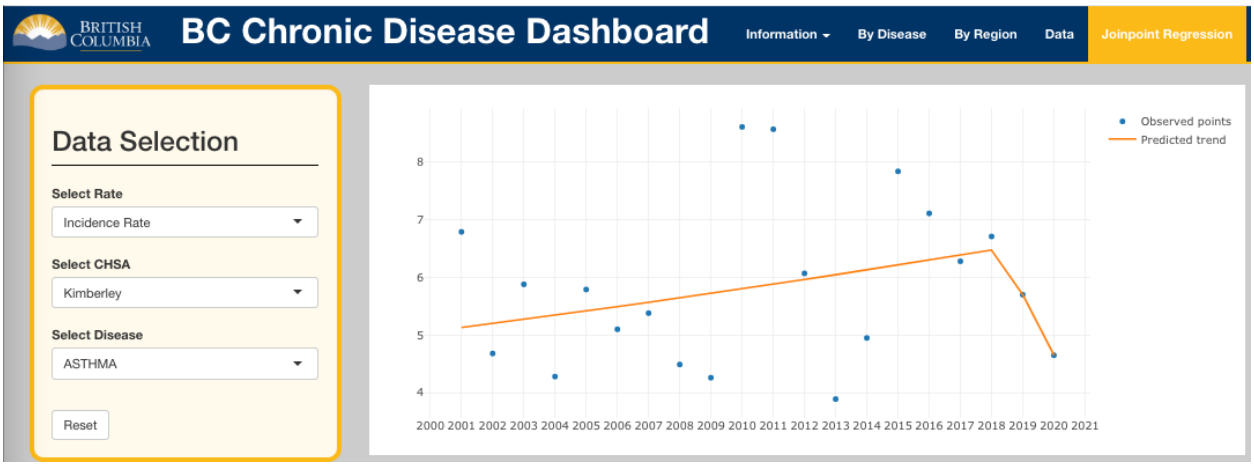

Figure 6: Layout of the 'Jointpoint Regression' tab of the dashboard

5 Conclusion and Recommendations

Every year, our capstone partner, the Office of British Columbia Provincial Health Officer, analyzes the interactions British Columbians have with the BC healthcare system and creates a data product called the Chronic Disease Registry that matches the healthcare utilization patterns to the case definition for 25 chronic conditions. They then aggregate these diseases' counts and rates by time, location, and sex. They sought for two well-defined deliverables: 1) an interactive data visualization tool using Shiny for the Chronic Disease Registry and 2) spatio-temporal analysis tools to interpret trends in the data.

After working diligently for seven weeks, our team of four successfully produce the two required data products. First, we build a Shiny dashboard that allows health professionals and non-experts alike to explore the distribution of diseases. The dashboard is fully interactive, user-friendly in that it provides the necessary information to navigate, and visually appealing and consistent with the BC government's color scheme. We implement all the proposed features, such as the four-tab layout design, showing a map alongside the graphs, and animating the plots to display yearly changes. Moreover, we add functionality and interactivity that ease and enrich the user experience: such as the hover information, linked highlighting effect for all the plots in one tab, and choosing whether y-axis should start from 0. We also use various methods to improve the app's performance for first loading and later rerendering.

Second, we generate models to interpret the temporal trends. We use a Bayesian modeling approach as first proposed to fit on 20-year data for each disease, CHSA, and age-standardized rate type. The fitted values smoothen the large fluctuations and thus wide confidence intervals in smaller health communities. However, 9% of the data, which are rare diseases with very low rates, can not be modeled using the Bayesian approach since the Gamma distribution does not provide support for zeros. We explore the local polynomial regression (LOESS) model as an alternative. The smoothened temporal trends are incorporated into the dashboard for users to easily compare trends between different CHSAs.

Last but not least, as something extra, we apply join-point regression to estimate abrupt changes in rates. The model can help epidemiologists locate the break points at a certain time and later draw inferences.

While we attempted to perfect the deliverables based on constant partner feedback and further research, due to the short time frame of this capstone project, we acknowledge that there are areas we can improve on. Here are the recommendations on future endeavors. For the dashboard, the main improvement lies in the processing speed. Currently, we have only tested the dashboard locally for a single user per session. The performance in terms of processing speed is acceptable, but varies greatly by individual laptops. If it is foreseeable that multiple users will be active simultaneously, asynchronous programming can be applied to increase scalability. Storing pre-processed data on a cloud server can be considered to decrease the start

up time. For more details on all our attempts to speed up the Shiny dashboards, please refer to this Github issue in the project repository.

To improve the Bayesian approach, additional fine-tuning of the model priors and scaling of the random walk model with the modified zero Gamma model is recommended.

References

Betts, Matthew G, Graham J Forbes, and Antony W Diamond. 2007. "Thresholds in Songbird Occurrence in Relation to Landscape Structure." *Conservation Biology* 21 (4): 1046–58.

Kurz, C. F. 2017. "Tweedie Distributions for Fitting Semicontinuous Health Care Utilization Cost Data." *BMC Med Res Methodol* 17 (171).

Muggeo, Vito MR et al. 2008. "Segmented: An r Package to Fit Regression Models with Broken-Line Relationships." *R News* 8 (1): 20–25.

RJ, Hastie TJ Tibshirani. 1990. "Generalized Additive Models." *CRC Monographs on Statistics & Applied Probability. New York: Chapman & Hall.*

Ulm, Kurt. 1991. "A Statistical Method for Assessing a Threshold in Epidemiological Studies." *Statistics in Medicine* 10 (3): 341–49.

Wang, Yue, X. 2018. *Bayesian Regression Modeling with INLA*. Chapman & Hall/CRC Press.