Analysis of Temporal Trends in Cancer Registry Data in NHS Scotland

Contents

Introduction

1.1 Background

Understanding trends within a countries cancer registry data provided an essential understanding for health budget allocation and informed public health polices, targeted at reducing the incidence of certain cancers. Cancer diagnoses places significant strain on not only a patients life, but on the health care budget allocated. This form of stress on an economies healthcare budget is known as the cancer burden. The larger the cancer burden the harder the economy must work to meet the growing demands placed on the healthcare system to provide the patients with the intensive care required. To illustrate this problem we can examine how , Europe currently has the highest health expenditure on cancer treatment also has the highest rate of deaths in a working age (30-69) due to cancer. Despite increased funding, quality healthcare and developing research into improving cancer care, providing the correct care in a timely manner that ensures cancer is caught earlier to prevent increased costs on a health care system still proves an issue. This is where maintaining and updating health statistics is important to inform policy makers on the changing needs amid rising cases.

Methods

Data Loading/Cleansing

First the relevant library packages are installed, there are som typical boilerplate packages. Of note to this paticular report, are the packages:

- Lemon - provides functionalities for geom_pointline
- kableExtra - provides functionlities to make a table in R Markdown
- Zoo - provides mathematical functions
- ggrepel - provides functionalities to ensure that text labels dont overalp on graphs

Load in Specific Packages

```
suppressWarnings({
library(readr)
library(tidyr)
library(dplyr)
library(here)
library(lemon)
library(kableExtra)
library(ggplot2)
library(data.table)

library(zoo)
library(ggrepel)
})
```

Once all the relevant libraries are loaded in, the data is read in from a local directory and saved into a dataframe. In this case the dataframe that will be used for all data manipulation of the entire cancer registry data will be named *cancerReg*

Load in the data

```
# Load in data
cancerReg <- read.csv("C:\\Users\\romin\\ToyRepo\\Models\\cancerReg.csv")

# Split the data, so structure can be seen better
cancerRegTable1 <- cancerReg[c(1,2,3,4,5,8)]
cancerRegTable2 <- cancerReg[c(6, 9, 10, 11)]

kable(head(cancerRegTable1))
```

area_code	area_type	area_name	year	period	numerator
S08000015	Health board	NHS Ayrshire & Arran	2003	2002 to 2004 calendar years; 3-year aggregates	2122.7
S08000016	Health board	NHS Borders	2003	2002 to 2004 calendar years; 3-year aggregates	650.0
S08000017	Health board	NHS Dumfries & Galloway	2003	2002 to 2004 calendar years; 3-year aggregates	916.0
S08000019	Health board	NHS Forth Valley	2003	2002 to 2004 calendar years; 3-year aggregates	1504.7
S08000020	Health board	NHS Grampian	2003	2002 to 2004 calendar years; 3-year aggregates	2651.7
S08000022	Health board	NHS Highland	2003	2002 to 2004 calendar years; 3-year aggregates	1661.7

```
#Continuation of table 1, too many columns to display in one table
kable(head(cancerRegTable2))
```

type_definition	measure	upper_confidence_interval	lower_confidence_interval
Age-sex standardised rate per 100,000	649.7	679.2	621.2
Age-sex standardised rate per 100,000	614.6	665.4	566.6
Age-sex standardised rate per 100,000	620.5	663.4	579.7
Age-sex standardised rate per 100,000	651.8	687.3	617.7
Age-sex standardised rate per 100,000	617.8	642.9	593.5
Age-sex standardised rate per 100,000	602.9	633.8	573.0

Not all the data provided in the initial data frame is required for further analysis and this line of code depicts the values in which will not be needed. These values have been removed to make the data cleaner and easier to manipulate. A notable exclusion is the “numerator” column. This column is the raw value of the amount of people registered under that specific health board in the cancer registry. This value could not be used as a reliable indicator of the rate of patients in the cancer registry by a specific health board, due to its lack of standardisation.

Standardisation is essential when looking at larger population data, especially with an aging population such as Scotland’s. Certain disease such as cancers are more prevalent in an aging population as seen with (insert reference). This is a significant factor in this study as Scotland is made up of several rural areas, in which their age distribution is significantly higher towards an elderly population. Specifically, between the more rural health boards such as NHS Orkney, NHS Ayrshire compares against more urban health boards such as NHS Greater Glasgow and Clyde and NHS Grampian. The biggest gap in age different is between Glasgow City and Dumfries and Galloway with a 14% to 27% difference respectively. Hence the importance of standardising the numerator value across the health boards, into a variable such as the measure column.

Remove Uncessary Data for Analysis

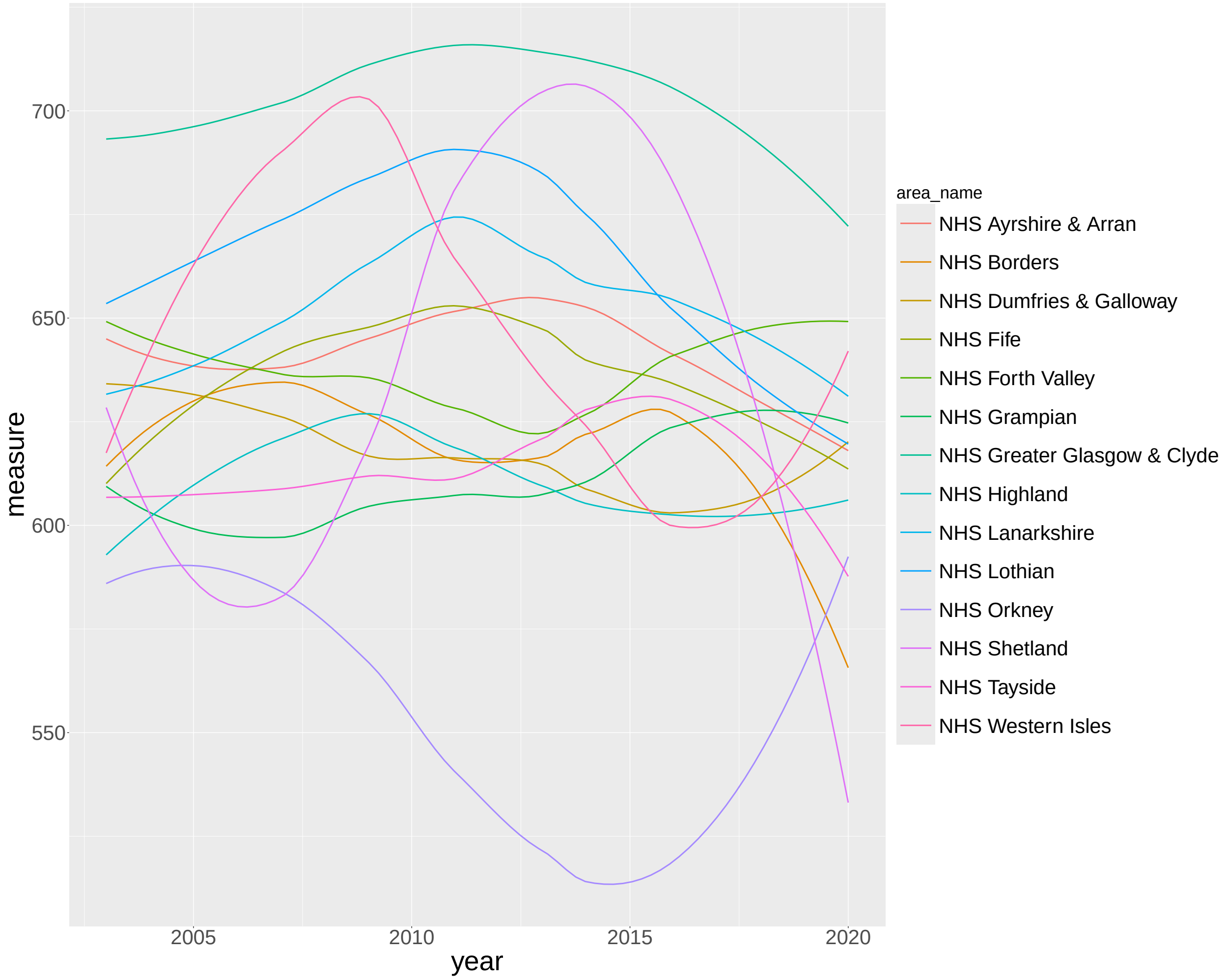
```r
cancerReg <- cancerReg %>% select(
    -period, -area_type, -type_definition, -indicator,
    -upper_confidence_interval, -lower_confidence_interval, -numerator
)
```

Visual Representation of All Values in Cancer Registry Data

To understand how to best analyse the data, we first require an understanding of what the data represents. Each health board every year submitted a raw value in the form of the numerator and the measure was then calculated against an age-sex standardisation method. After the data was adjusted, the differences in measure values among health boards were minimal, therefore there was a large amount of health boards falling into the same value.This caused a major issue with overplotting across different health boards. To try and overcome the issue of overplotting, a variety of visual graphs were tested. Such as density plots, heatmaps and stacked area charts, but they all proved too limited in showing the drastic jumps some health boards exhibited in certain time frames.

This portion of the assignment took the longest due to the difficulty of clearly mapping all 14 health boards over an extended period. Especially when there are such drastic changes in the measure, often these sharp changes in the measure value were lost in favour of generalising the data. Therefore, the decision was made to use a scatter and line plot with varying functions designed to smoothen the lines between certain points where they seemingly overlap consistently.

Display All Data Points



The function *geom_smooth()* was used to account for the overplotting that occurred for the majority of the health boards in their respective time periods. This function was particularly useful as it creates smoother transition lines between the physical dots representing the measure value across the years. For the overall simplicity of the graph, the physical dots representing the data have been omitted, to ensure a clear overview of the data.

The data itself shows clear jumps between measure values between the health boards. To better understand the scale by which the measure values differ, an overall mean will be calculated for all values, while a cummulative moving average will be used to calculate the avergae of a specific health board.The aim is to quantify how a health board's cancer registry data differs from the overall average from 2003-2020. To analyse this relationship, the group moving average is calculated against an individual health board. In calculating this, we can numerically evaluate how the values have changed. While also investigating if a particular year showed an increase/decrease of patients in the cancer registry.

Calculate Overall and Moving Average for All Health Boards

Find Average of All Measures by Year

```
# Total Average across all health boards, grouped by the year
avgYearly <- cancerReg %>%
    group_by(year) %>%
    #Add a row showing the average number for the measure from all healthboards in a year
    mutate(AvgYear = mean(measure, na.rm = TRUE)) %>%
    select(-area_name, -measure, -area_code)
```

Since this report focuses on time series data, it would be appropriate to use time series methods to analyse the effects over the years accurately. Given the vast amount of historical data provided, it was important to consider how to incorporate the historical data into the analysis. While most moving average models place emphasis and weightings on more recent data points, the cumulative moving average considers most historical data and provides the means of examining long term trends for patients within the cancer registry.

For the purposes of this report, we aim to explore fluctuations in registry data over the years, particularly how the pandemic has impacted the cancer registry. We are interested in observing the increase/decrease in cancer trends to then explore if the pandemic did cause an effect on registry data, or if there were underlying trends in previous years that have exacerbated the status quo. Therefore, the decision was made to use the cumulative average, to maintain the weighting that historical data would have on the overall value, so that a clearer distinction could be made on each year's trend in cancer registry data.

Calculate Moving Average for Each Health Board

```
# Calculate a cummulate moving average for individual healthboards
movingAvg <- cancerReg %>%
    group_by(area_name) %>%
    arrange(year) %>%
    mutate(MA = cumsum(measure) / row_number())
```

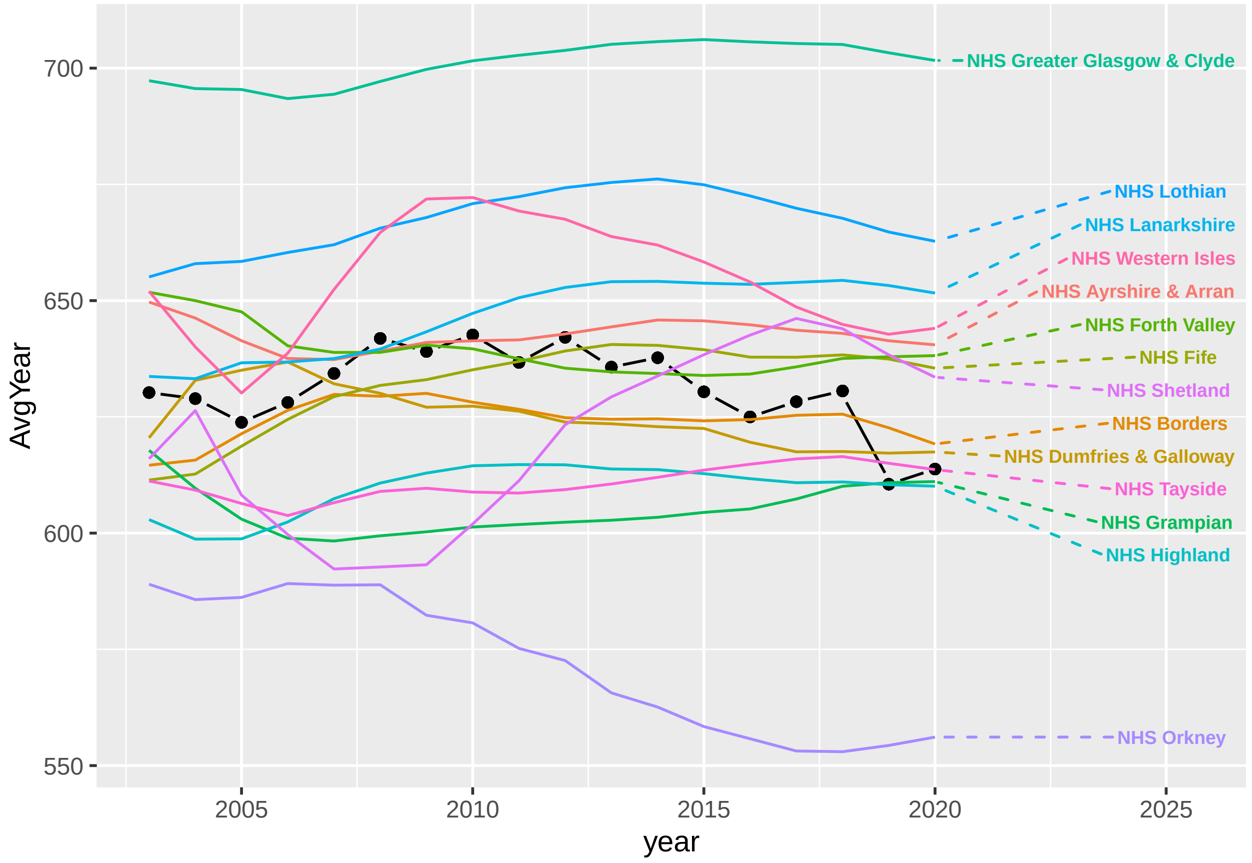
Find Last Data Points for Data

```
#Used for asethetic purposes only
#Code used to assign easier to read labels on graph
finalValues <- movingAvg %>%
    group_by(area_name) %>%
    summarise(
        lastMA = dplyr::last(MA),
        lastYear=dplyr::last(year)
    )
```

Display Summary of All Data

```
    options(repr.plot.width = 60, repr.plot.height =5)
    ggplot(data = cancerReg, aes(x = year)) +
        geom_pointline(data = avgYearly, aes(y = AvgYear)) +
        geom_line(data = movingAvg, aes(y = MA, col = area_name)) +
        geom_text_repel(
            data = finalValues, aes(
                x = lastYear,
                y = lastMA,
                label = area_name,
                color = area_name
            ),
            size = 2.5,
            fontface = "bold",
            nudge_x = 5.6,
            direction = "y",
            hjust = 0.7,
            segment.linetype = 2,
            segment.size = 0.5,
            segment.curvature = 0,
            max.overlaps = Inf
        ) +

        theme(legend.position = "none", plot.margin = margin(2,2,2,2))
```



The overall average is calculated and plotted as a scatter plot over a line plot of the individual health board averages. The lines that were previously plotted on the overview of the data have been smoothed to show the uptick in trends over time. Notably, there are few health boards that have consistently remained within the overall average, most of these are contained within the middle of the labels. As the labels of the graph move further out, we see stronger variations in the measure value over time. Specifically, NHS Western Isles, NHS Orkney and NHS Shetland. These health boards consistently have shown varying levels of fluctuation over the years. Without using the cumulative summation to calculate these averages, their historical data would have been lost.

The moving average itself saw a slight uptick in cases during 2006-2008 and the cases yo-yoed up and down for the next 7 years until there was a sharp decline. This decline shows that during 2019, across all health boards the measure of patients in the cancer registry fell. To quantify this value we can calculate, the differences between the cumulative moving average and the overall average. The difference provided would show the rise (positive difference) or fall (negative difference) in patients in the cancer registry, against the national average. Providing a clear indicator of trends over time, to see if the pandemic attributed to a change in cancer registry data, or if there was already a pattern of change before the pandemic.

Quantify Differences Between the Averages

Calculate Differences Function

```r
#Declare data frame to hold percentage values
sigPercent <- data.frame(
    area_name = character(),
    year = integer(),
    percentNum = numeric(),
    stringsAsFactors = FALSE
)
boardAvg <- function(currBoard, currVal, currYear) {
    #Retrive current average for a specific year
    currAvgYear <- filter(movingAvg, area_name == currBoard & year == currYear) %>% select(MA)
    numCurrAvgYear <- gsub("[^0-9.]", "", currAvgYear$MA)
    numCurrAvgYear <- as.numeric(numCurrAvgYear)
    #Calculate difference
    diffVal <- currVal - numCurrAvgYear
    percentVal <- ((diffVal / numCurrAvgYear) * 100)
    # Defined a threshold, where the difference percentage would be of interest
    if (percentVal >= 3 || percentVal <= -3) {
        sigPercent <- sigPercent %>% add_row(
            area_name = currBoard,
            year = currYear,
            percentNum = round(percentVal, 2)
        )
    }
    return(sigPercent)


}
```

While the moving average does not provide a direct estimation of the predicted values, it still serves as an indicator of what the value would have looked like in that time frame. A large deviation from that number means there was a change in the previous trend of data. To investigate such a change, we can calculate the difference in the value. Since the cumulative summation considered historical data, it can provide an estimation of how different that value is from the overall trend. In this case the trend is the overall average. It would be typical to see the value of the measure fluctuate slightly over time and therefore this difference would not be considered as a cause for interest.

A value had to be determined to capture meaningful changes in the data. The difference was expressed as a percentage, with 100% indicating that the current point difference matched the average cases for that year, and 0% indicated no deviation from the point. Looking at the graph, meaningful deviations meant any case that was ± 30 cases, this translates to a 3% change from the overall average value. 3% was used as the threshold, any health board for a specific year that encountered a change of at least 3% of the overall average was saved into another data frame, to have the contents further analysed.

Calculate Differences

```r
healthBoards <- unique(cancerReg$area_name)
totalYears <- unique(cancerReg$year)
for (currBoard in healthBoards) {
    for (currYear in totalYears) {

        currVal <- subset(cancerReg, year == currYear & area_name == currBoard)
```

```
            currVal <- select(currVal, -area_code, -area_name, -year)
            currVal <- as.numeric(currVal)
            #Call difference function
            sigPercent <- boardAvg(currBoard, currVal, currYear)
        }
    }
    sigPercent <- sigPercent %>% arrange(desc(year))
```

##Disp

```
sigPercentWide <- sigPercent %>% pivot_wider(
    names_from = year,
    values_from = percentNum
)

sigPercentWide <- sigPercentWide %>% replace(is.na(.), 0)

# Add reference to statology
sigPercentTotal <- sigPercentWide %>%
    bind_rows(summarise(
        ., across(where(is.numeric), sum),
        across(where(is.character), ~"Total")
    ))

kable(sigPercentTotal, format = "latex", booktabs = TRUE, longtable = TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_position")) %>%
    row_spec(0, bold = TRUE) %>%
    column_spec(1, "2cm") %>%

    row_spec(15, bold = TRUE) %>%
    kableExtra::landscape()
```

area_name	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005
NHS Borders	-9.54	-7.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NHS Lothian	-5.11	-7.11	-4.85	-5.55	-4.63	0.00	0.00	0.00	0.00	0.00	3.12	0.00	0.00	0.00	0.00	0.00
NHS Orkney	5.48	3.89	0.00	-6.66	-6.13	-9.02	-5.95	-12.29	-4.10	-7.63	0.00	-6.75	0.00	0.00	0.00	0.00
NHS Shetland	-12.88	-14.09	-5.15	7.74	8.53	8.62	7.78	9.53	17.39	12.16	10.21	0.00	0.00	-5.01	-4.24	-5.98
NHS Western Isles	3.38	-5.25	-8.71	-11.61	-8.55	-6.64	-3.04	-5.60	0.00	-3.48	0.00	6.42	9.23	8.40	4.02	-3.15
NHS Fife	-5.27	0.00	0.00	0.00	-3.25	0.00	0.00	0.00	3.17	0.00	0.00	0.00	0.00	3.11	0.00	0.00
NHS Tayside	-3.77	-3.74	0.00	0.00	0.00	3.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NHS Greater Glasgow & Clyde	-4.01	-4.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NHS Lanarkshire	-4.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.04	4.13	4.29	3.47	0.00	0.00	0.00	0.00
NHS Ayrshire & Arran	0.00	-3.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NHS Forth Valley	0.00	0.00	4.26	3.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-3.43	0.00
NHS Grampian	0.00	0.00	6.72	4.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NHS Dumfries & Galloway	0.00	0.00	0.00	-4.66	-6.24	0.00	0.00	0.00	-3.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NHS Highland	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.29	0.00	0.00
Total	**-35.94**	**-41.87**	**-7.73**	**-12.39**	**-20.27**	**-4.02**	**-1.21**	**-8.36**	**16.11**	**5.18**	**17.62**	**3.14**	**9.23**	**9.79**	**-3.65**	**-9.13**

In line with the previous graph of the data points, we see that the drop in the graph was between 2019 and 2020. During those time periods 9/14 health boards saw a decrease in the overall number of patients in their respective cancer registries. It's important to note at this point that the data collected come from 3-year aggregates. Any data provided from a specific year also includes data from the previous years. Therefore, the 2019 data is part of a 3-year aggregate containig data from 2018-2020. This form of collecting data as an aggregate was unclear. In the definition and dictionary of the data provided on the PHS website, it states that to accurately represent the true value of a year's cancer registry data they require the next years data. This purpose was unclear. In light of this information, we can assume that the data from 2019 and 2020 come from a time during the pandemic and further analyse what the percentage changes mean for each of the health boards.

The pandemic showed a clear change amongst all health boards, in addition to another year of interest in 2016 that say a total of 20% cut in cancer registry data. 2016 was a notable year for Scotlands economy and caused a shift in Scotlands work force, this change will be further discussed in the Results section.

There seems to be a consistent number of health boards that contribute to these changes, especially NHS Western Isles, NHS Orkney and NHS Shetland. These health boards experienced a deviation almost every year, this doesn't necessarily indicate a worsening state, as lower cancer rates will also be flagged as a change from the normal. If a specific health board continuously has changes from the average in a negative direction it could indicate a healthier population.