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)
```

```r
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

```r
cancerReg <- read.csv("C:\\Users\\romin\\ToyRepo\\Models\\cancerReg.csv")
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

```r
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 inital dataframe is required for further analysis and this line of code depicts the values in which we will not need. 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 under a specific health board in a specific time frame are contained in the cancer registry. This was due to the fact that the “numerator” values were not standardised across different health boards, and if the raw values were taken, it could lead to false conclusions over the data.

Especially with the variance in population age between the healthboards, the largest difference being between Dumfries and Galloway with over 65’s representing 29% of their population compared to Grampian’s 20%. As noted in the Background section most new incidence of cancer come from an aging populaion. So for the sake of genralisability of results, a standardised value such as the measure is used throughout the interperetations and calulcations. The measure value itself is not standardised to a fine degree, but this will be further discussed in the results section

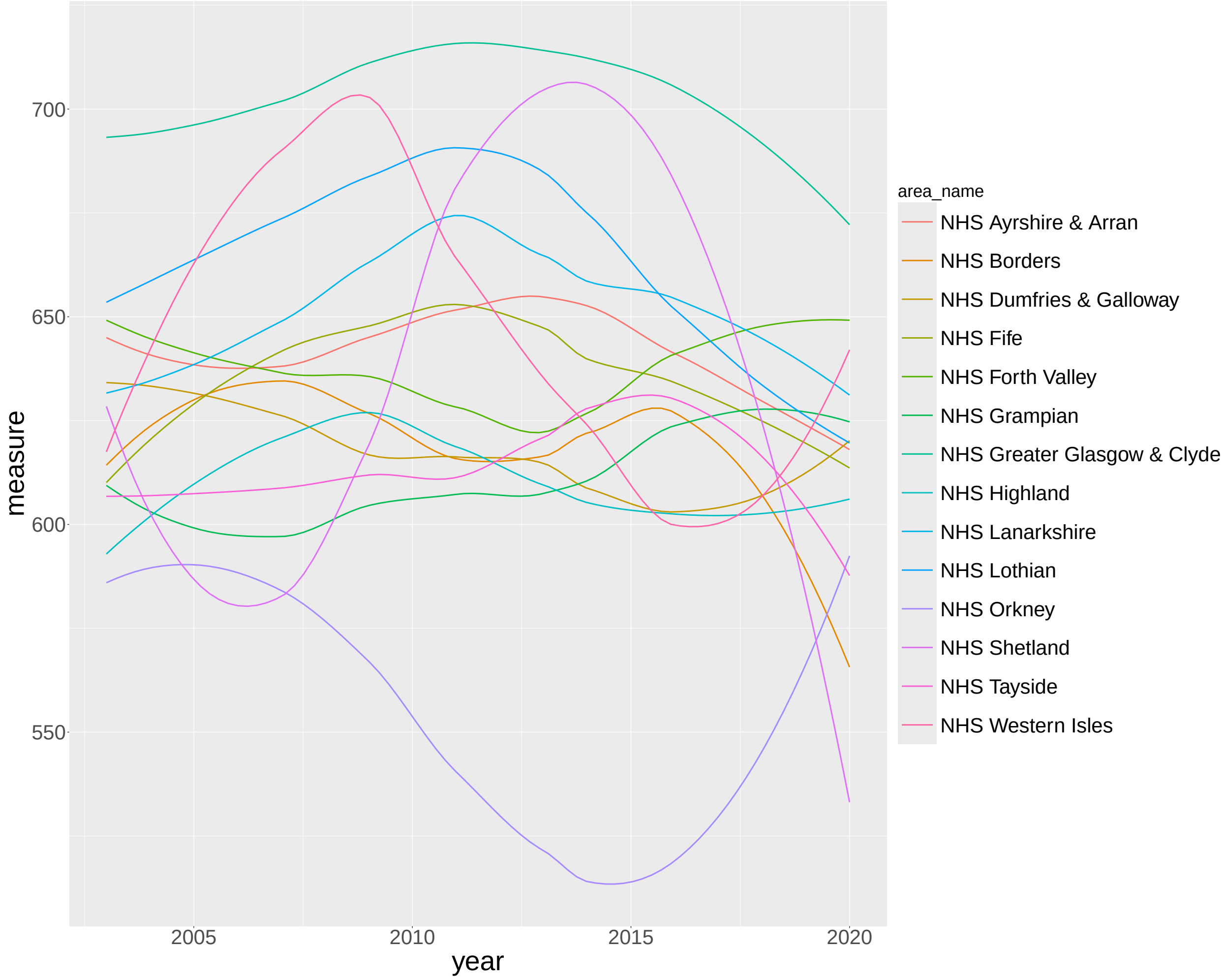
Remove Uncessary Data for Analysis

```
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, a general overview is helpful to aid in this. This portion of the assignemnt took the longest since it was difficult to map for all 14 healthboards in a clear and conscise manner without cluttering the plot. especially when there such drastic changes in the measure variable between time periods. Initally, a variation of plots such as stacked area chart and density plots were experimented with but they all proved too limited in their visual represnations as overlapping data, and the jumping of values were often lost in favour of generalising the data points. Therefore, the descion was made to use a scatter plot with lines intersecting the points to display the change in gradient of the line for different time periods.

Display All Data Points



Find Average of All Measures by Year

```
avgYearly <- cancerReg %>%
    group_by(year) %>%
    mutate(AvgYear = mean(measure, na.rm = TRUE)) %>%
    select(-area_name, -measure, -area_code)
```

Calculate Moving Average for Each Health Board

```
movingAvg <- cancerReg %>%
    group_by(area_name) %>%
    arrange(year) %>%
    mutate(MA = cumsum(measure) / row_number())
```

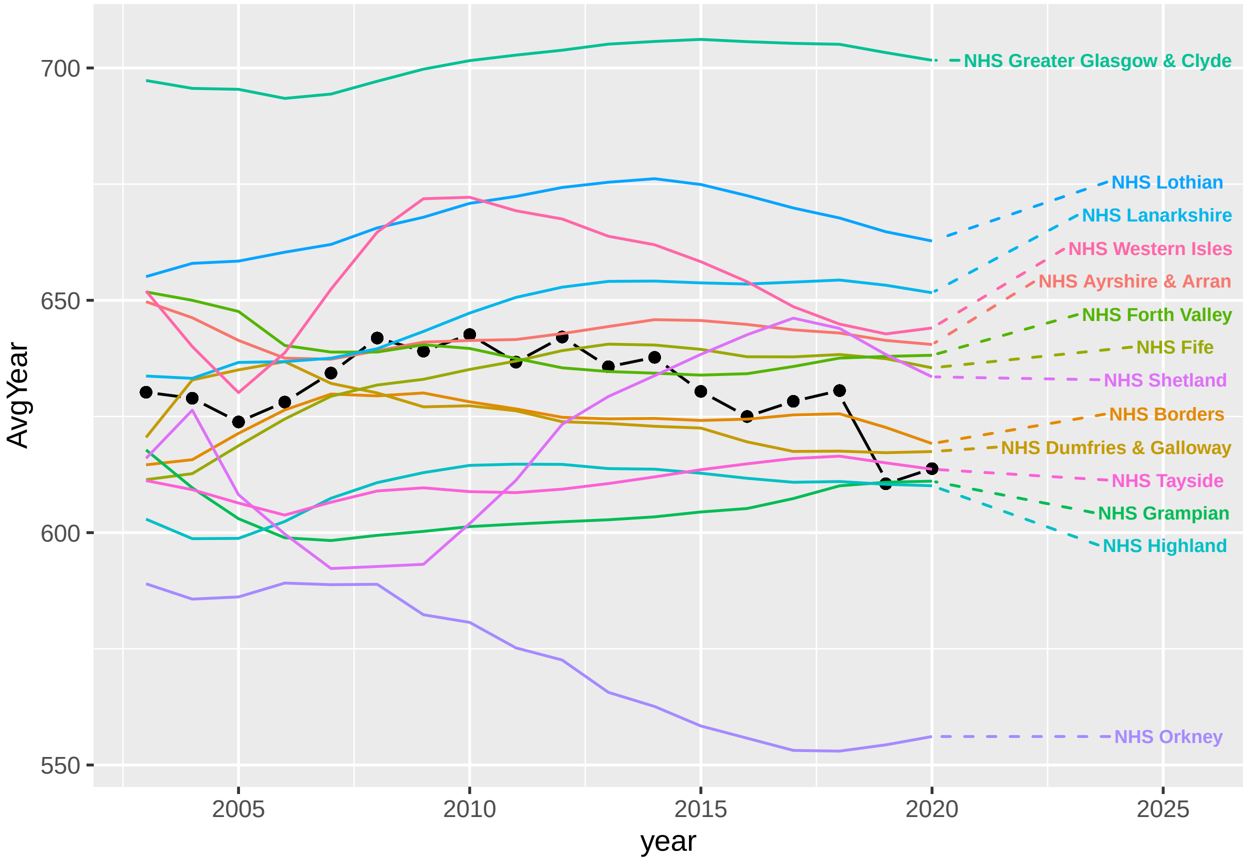
Find Last Data Points for Data

```
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))
```



Calculate Differences Function

```
sigPercent <- data.frame(
    area_name = character(),
    year = integer(),
    percentNum = numeric(),
    stringsAsFactors = FALSE
)
boardAvg <- function(currBoard, currVal, currYear) {
    currAvgYear <- filter(movingAvg, area_name == currBoard & year == currYear) %>% select(MA)
    numCurrAvgYear <- gsub("[^0-9.]", "", currAvgYear$MA)
    numCurrAvgYear <- as.numeric(numCurrAvgYear)
    diffVal <- currVal - numCurrAvgYear
    percentVal <- ((diffVal / numCurrAvgYear) * 100)

    if (percentVal >= 3 || percentVal <= -3) {
        sigPercent <- sigPercent %>% add_row(area_name = currBoard, year = currYear, percentNum = r
    }
    return(sigPercent)

    # WHile the moving average does not provide a direct estimation of the predictied values it sti
}
```

Calculate Differences

```
    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)
            sigPercent <- boardAvg(currBoard, currVal, currYear)
        }
    }
    sigPercent <- sigPercent %>% arrange(desc(year))
```

bjkkk

```
# inputFile <- "reportReg.pdf"
# sigPercentWide_colored <- sigPercentWide %>%
#     mutate(across(everything(), ~ cell_spec(.,
#                                              color = ifelse(. < 0, "red", "black"),
#                                              background = ifelse(. < 0, "lightpink", "white"))))


sigPercentWide <- sigPercent %>% pivot_wider(
    names_from = year,
    values_from = percentNum
)
# sigPercentWide <- sigPercentWide %>% mutate(
#   across(-1,

#       ~ cell_spec(., color = ifelse(is.na(.), "black", ifelse(. < 0, "green", "red")))
#   )
#   )
# sigPercentWide <- sigPercentWide %>% mutate(across(everything(), ~ replace_na(., 0)))
kable(sigPercentWide, format = "latex", booktabs = TRUE, longtable=TRUE) %>%
    kable_styling(latex_options = c("striped", "hold_postion")) %>%
    row_spec(0, 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	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NHS Lothian	-5.11	-7.11	-4.85	-5.55	-4.63	NA	NA	NA	NA	NA	3.12	NA	NA	NA	NA	NA
NHS Orkney	5.48	3.89	NA	-6.66	-6.13	-9.02	-5.95	-12.29	-4.10	-7.63	NA	-6.75	NA	NA	NA	NA
NHS Shetland	-12.88	-14.09	-5.15	7.74	8.53	8.62	7.78	9.53	17.39	12.16	10.21	NA	NA	-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	NA	-3.48	NA	6.42	9.23	8.40	4.02	-3.15
NHS Fife	-5.27	NA	NA	NA	-3.25	NA	NA	NA	3.17	NA	NA	NA	NA	3.11	NA	NA
NHS Tayside	-3.77	-3.74	NA	NA	NA	3.02	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NHS Greater Glasgow & Clyde	-4.01	-4.05	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NHS Lanarkshire	-4.22	NA	NA	NA	NA	NA	NA	NA	3.04	4.13	4.29	3.47	NA	NA	NA	NA
NHS Ayrshire & Arran	NA	-3.96	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NHS Forth Valley	NA	NA	4.26	3.40	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	-3.43	NA
NHS Grampian	NA	NA	6.72	4.95	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NHS Dumfries & Galloway	NA	NA	NA	-4.66	-6.24	NA	NA	NA	-3.39	NA	NA	NA	NA	NA	NA	NA
NHS Highland	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.29	NA	NA

```
# qpdf::pdf_rotate_pages(inputFile, pages = 4, angle = 90)
```

#Note for next time: what I want to do at this point is to show the changing colours as a difference change if its only within a small amount of chaning values then ignore the calues and do not #colour the cell, otherwise red fir a rise and green for a fall #