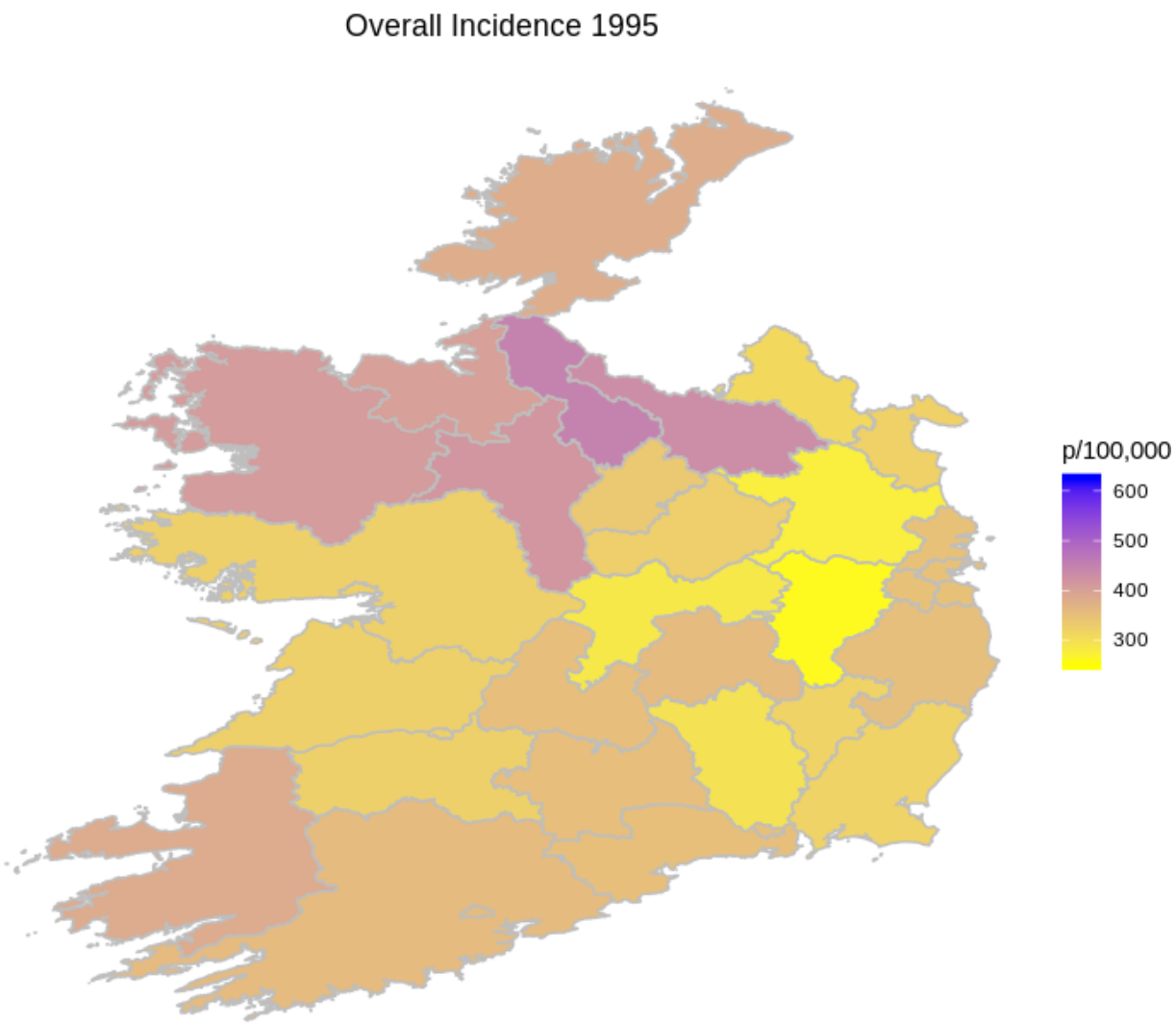
Data Visualisation & Analytics Project

Higher Diploma in
Data Science & Analytics
Liam De Barra




## Introduction

This project is an effort to summarise invasive cancer incidence in the Republic of Ireland from 1994 until 2015 using data collected by the National Cancer Registry of Ireland [1]. Some important caveats while interpreting cancer incidence data are the advent of better methods of detection (PSA screening for example) and an aging population which will inherently have a higher incidence of cancer. The focus of this study was incidence, and mortality of described cancers was not considered. It is the opinion of the author that mortality rates reported are inherently inaccurate as they are reported in terms of '5 year survival'. Thus, mortality can appear to be decreasing sharply when in fact cancers are simply being detected earlier; clouding the interpretation of 5 year survival rates. Crude incidence rates across all ages were used because childhood cancer rates are extremely low and unlikely to distort the analysis [2]. All rates reported do however exclude non-melatonomous skin cancer as is common practice due to it's low mortality and non-invasive pathology.

## Exploratory Data Analysis

Crude incidence rates are reported in ‘cases per 100,000’ (hereby shorthanded as CPO) and these data were collected for combined incidence of all cancers as well as for individual cancers of interest. Rates reported in CPO have the potential to mislead due to widespread reporting of cancer risk as ‘1 in 3’ or ‘1 in 4’, and require context for proper interpretation. CPO refers to the number of new diagnoses in any given year; so a rate of 500 cases per 100000 would imply a risk of 1 in 200 (p = .005) to the individual. However, this is the average probability of developing cancer in a single year and is therefore cumulative as a person ages. Ignoring other confounding factors and multiplying this risk by a life expectancy of e.g. 80 years results in a ‘lifetime risk’ of 80 * .005 = .4 or “1 in 2.5” commonly reported in the media.

The first factor investigated in this study was the combined cancer incidence rate by sex. Incidence increased during the study period for both sexes with a 38.16% increase in males (357 to 493 CPO) and a 29.86% increase for females 333 to 433 CPO) for an overall average increase of 34.05%. Although outside the scope of this study; brief investigation of median age in Ireland during the same period indicated an increase from 30.1 to 36.9 years between 1995 and 2015 [3]. Age is one of the greatest risk factors of cancer and a major caveat in interpreting these data. Gender distribution was assumed to be approximately equal but was not confirmed. The cancer rates for men and women during the study period were visualised using a line graph (Figure 1) and a colour scheme of blue for males and red for females chosen which was adhered to for the rest of the analysis. Cancer incidence data was manipulated using Excel and R’s dplyr package extensively.

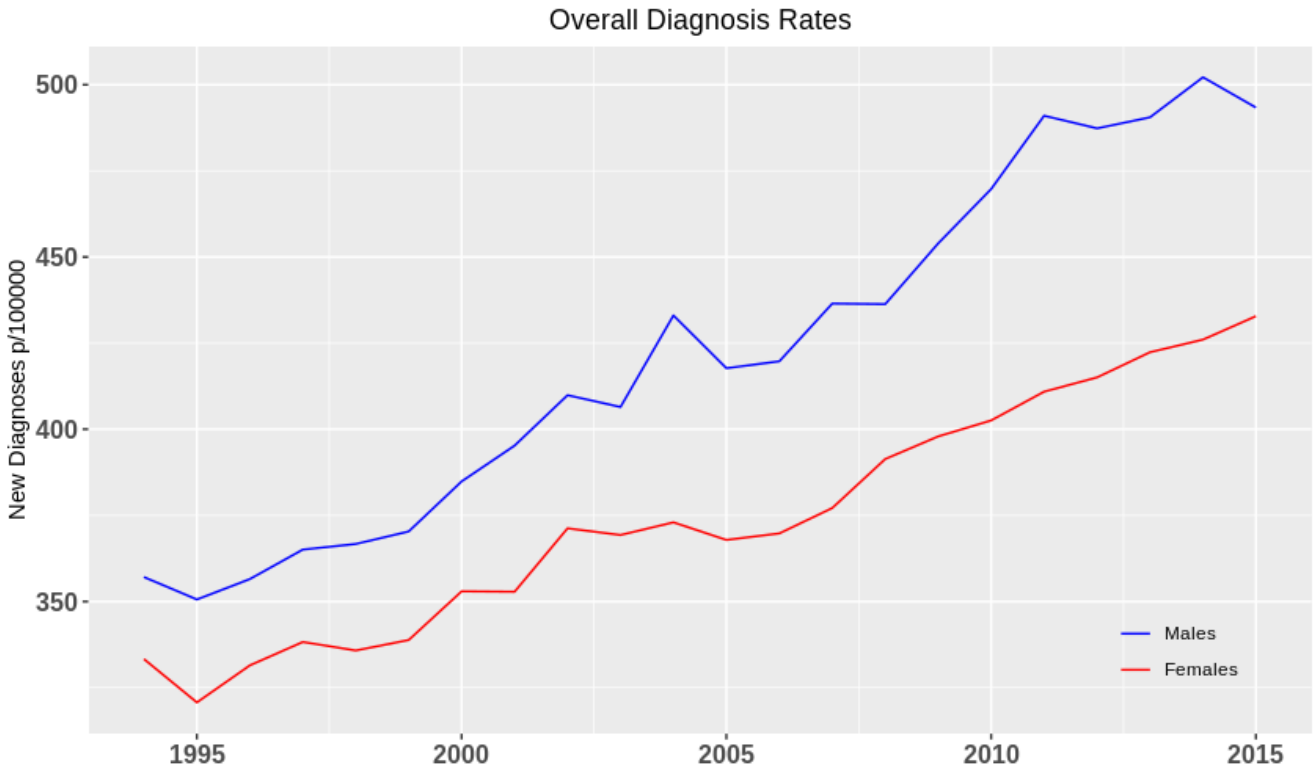



*Figure 1: Combined cancer incidence in males vs females*

**Most Common Cancers**

The most common cancers were next analysed in greater detail. The selection of 'most common' cancers however; turned out to be quite complex and in effect somewhat arbitrary. The most common cancer (non-melanomous skin cancer or NMSC) is generally excluded from cancer incidence figures due to it's low mortalty and the 5-most common cancers also underwent some minor changes during the study period. Furthermore, certain cancers of different tissues within an anatomic region are often reported together (e.g. gynaecological cancers), further complicating consistency in reporting. The five most common invasive cancers (excluding NMSC) for each sex at the end of the study period were chosen as suitable for the purpose of this project. Cancer of the corpus uteri was the 5$^{th}$ most common cancer among women and the single most common gynaecological cancer. Although it was potentially of interest to investigate incidence of cervical cancer in light of recent evidence of significant failings in Ireland's cervical cancer screening this was deemed to be beyond the scope of this project.

A horizontal bar chart was deemed the most appropriate depiction of these data as it clearly shows the inordinate contribution of breast & prostate cancer to overall figures. It also correctly implies to the viewer that despite these cancers being virtually unique to one gender; they still emerge as the two most prevalent cancers in terms of overall incidence. In other words, prostate and breast cancer are the 1$^{st}$ and 2nd most common invasive cancers in Ireland despite approximately half the population being susceptible to either one. A back to back chart was chosen as it shows the oddly symmetric distribution of the most common cancers for both sexes while also conveying to the astute observer that cancer is more prevalent in males Figure 2. The 'plotrix' package was used despite not offering the functionality to append different labels to either side of the chart (this was completed manually using the image processing program Shutter). The fact that these top 5 account for 62.6% of total cancer in males and 62.1% in females justified focusing on just these cancers in the remainder of the analysis.

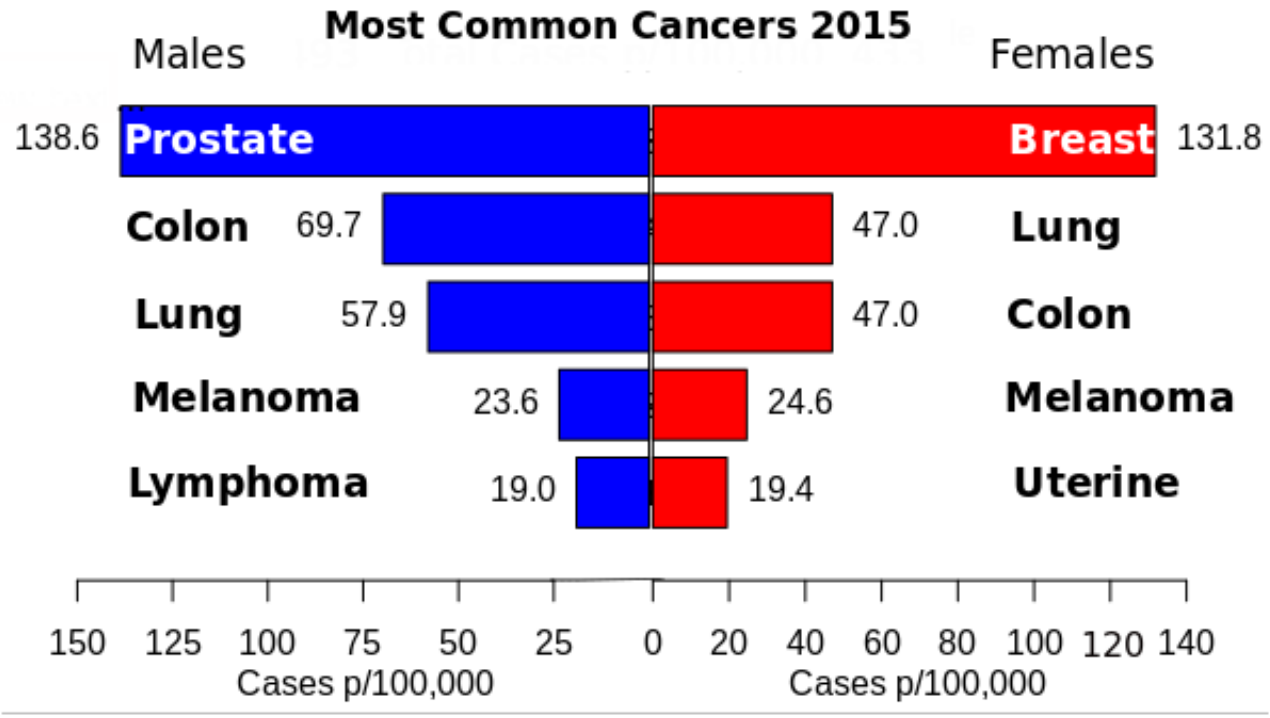



*Figure 2: Rates of 5 Most Common Cancers by Gender. Edited output from Plotrix.*

**Overall Incidence by County**

Prior to visualisation the changes in overall NMSC cancer during the study period were examined manually in excel using pivot tables and increased rates were observed in all 26 counties. Quite surprisingly however, the proportional increase in cancer incidence by county deviated drastically during the study period, from just 12% in Cavan to 61% in Louth; with a country-wide average of 34.05% (and standard deviation $\sigma = 11.7\%$). Tables of the 5 counties with largest & smallest incidence rates as well as those with the largest & smallest relative increase during the study period as output by dplyr and tabulated in Libre Office Impress are shown below. (Table 1 & 2). A pattern of higher incidence in the north-west of the country was observed which persisted throughout the study period. Although the cause was not investigated it was postulated that emigration among younger people may be particularly high as these regions are furthest from the major urban centers of Dublin and Cork.

| Largest Increase | Smallest Increase |
|---|---|
| Louth **61.1%** | Cavan **12.2%** |
| Offaly 54.3% | Laois 20.7% |
| Wexford 49.3% | Leitrim 22.7% |
| Carlow 48.8% | Longford 23.8% |
| Galway 48.7% | Dublin 24.5% |

*Table 1: Counties with largest & smallest relative increase in cancer rate (all cancers minus NMSC)*

| Highest Rate | Lowest Rate |
|---|---|
| Mayo: 600 | Meath 368 |
| Leitrim 551 | Kildare380 |
| Roscommon 543 | Kilkenny 396 |
| Kerry 534 | Longford 418 |
| Louth 521 | Laois 433 |

*Table 2: Counties with the highest & lowest cancer rates in 2015 (all cancers minus NMSC)*

## Overall Incidence Heatmap

Visualisation of overall incidence data utilised modified code sourced from a CSO housing price plot presentation [4] which used the “rgeos” and “maptools” libraries to construct Irish county borders from an spdf file. Because of discrepancies in defining borders (e.g. Cork County vs Cork city) and extraneous data such as HSE North/South-East the incidence data required extensive modification to visualise the incidence data. An overview of processing steps is provided below:

1. The overall national average and HSE jurisdiction incidence rates were removed and each county reduced to one overall figure (e.g. ‘Dublin’ replaced both Dulin North & Dublin South).
2. An excel spreadsheet with yearly cancer rates for the 26 counties was compared with the shape object used to plot housing data (which had a total of 32 areas including Tipperary North/South etc).
3. Incidence data for missing or sub-divided regions was imputed in the same fashion as step 1 to ensure number of rows matched; i.e. the cancer incidence value for ‘Cork’ was entered for both Cork city and Cork county.
4. Manually created incidence data columns were inserted into template code for creation of heatmap data.
5. A total of 32 data points were plotted from the data available for 26 counties but it should be pointed out that although certain counties are visibly sub-divided on the map; a single overall incidence figure was used in every case.

Data for 1995, 2005 and 2015 were chosen for a small multiples heatmap visualisation and optimisation centered around choosing an incidence range that highlighted the substantial increase in cancer rate without exaggerating or misleading the viewer (Figure 3). After step-wise testing & comparison; limits of 250 to 625 cases p/100000 were deemed to give the most effective and balanced output. Similar testing was carried out with graded colour schemes and depicting low rates in yellow vs high rates in purple emerged as the most striking colour scheme. Purple was purposefully chosen in the hope of reiterating to the viewer that these data were for male & females combined (mixture of blue & red). The heatmap captures the aforementioned pattern of higher incidence in the northwest as well as a pocket of low incidence in the east midlands (Meath). Enlarged heatmaps are included in appendix

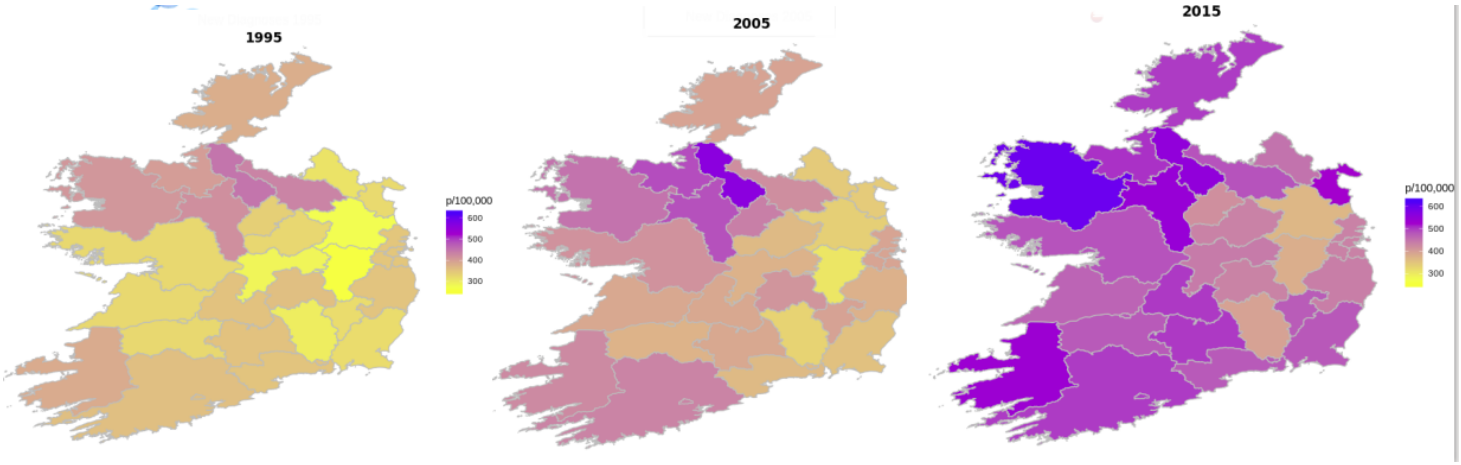



*Figure 3: Heatmap of total cancer incidence by county*

**Investigation of Relative Increase**

Incidence of the top 5 cancers was analysed in greater detail to assess if any cancer had a particualrly sharp increase in incidence during the study period (compared to the overall average of 34.05%). Required data was collected using dplyr and output to Excel for easier manipulation. The results showed a disproportionately large change in rates of several cancers with melanoma & prostate cancer in males seeing the largest increases (Table 3). Another surprising factor was the 3% decrease in lung cancer rates in males. While this appears small at first glance; when considered in the context of the average increased cancer rate in males of 38.1% it represents a substantial curbing of male lung cancer possibly attributable to the smoking ban and public health campaigns aimed at reducing smoking i the late 90s. Further investigation of this finding as well as comparison with female rates was not carried out due to time restraints but gave promising evidence for the potential success of implementing public health initiatives in reducing cancer rates.

| Relative Change 1994-2015 (μ = 1.34) | | | |
|---|---|---|---|
| 1.Melanoma (M) | 3.05 | 6.Lung (F) | 1.66 |
| 2.Prostate | 2.25 | 7.Breast | 1.52 |
| 3.Uterine | 1.92 | 8.Colon (M) | 1.21 |
| 4.Melanoma (F) | 1.78 | 9.Colon (F) | 1.05 |
| 5.Lymphoma (M) | 1.74 | 10.Lung (M) | 0.97 |

*Table 3. Relative changes in most common incidence rates between 1994 and 2015. National Average displayed in yellow for comparison with Figure 4*

Initial efforts to compare the distribution of the rates for the most common cancers between 1994 and 2015 using a boxplot lacked clarity and did not convey the desired message that male melanoma rates in particular had greatly increased. Instead, breast & prostate cancer appeared as outliers despite a comparatively small increase in breast cancer rate in the study period. This was due to to the large inter-cancer variance in rate and no suitable means of transforming these data could be found. The original boxplot output is included in the appendix but a simple boxplot of the relative increase in each cancer between 1994 & 2015 highlighted both high rate of increase (among what were already the most common cancers) and the profound increase in male melanoma (Figure 4).

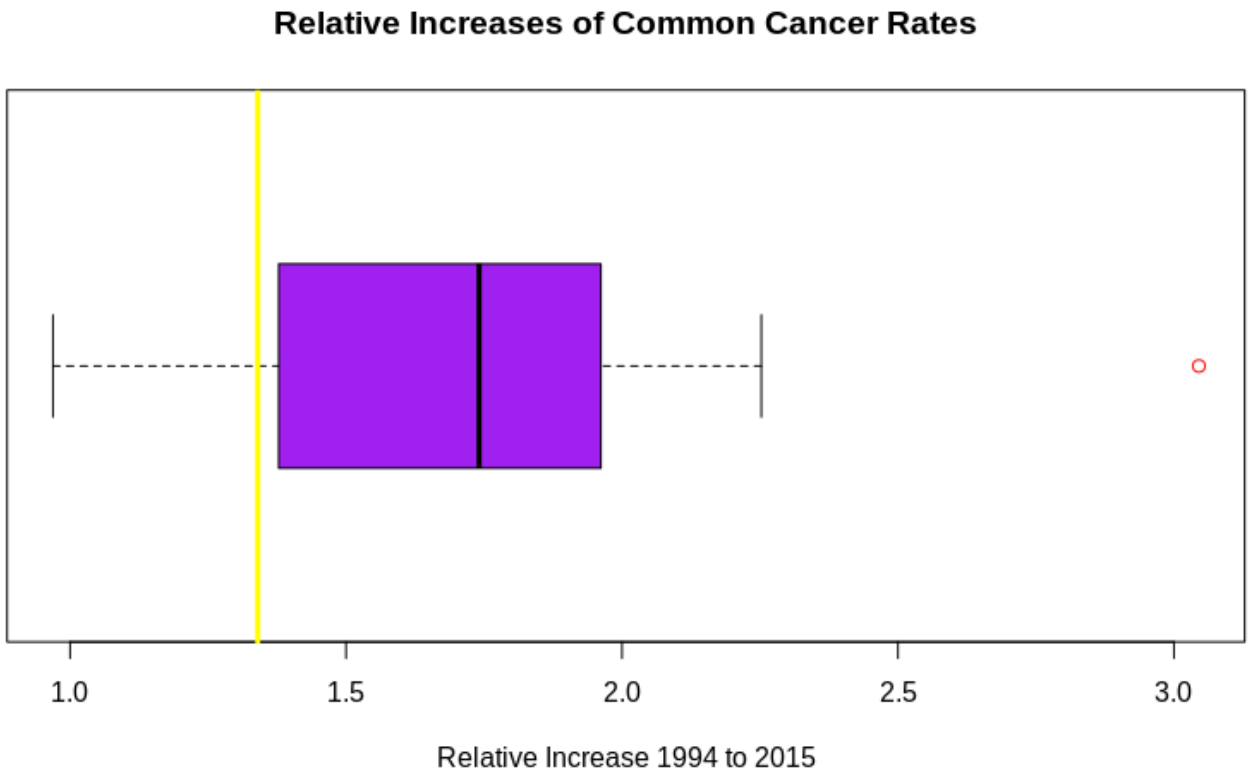



*Figure 4: Distribution of relative increases for the top 5 cancers in both sexes. Overall average increase across all cancers shown in yellow with male melanoma shown in red.*

**Male Melanoma**

The exceptionally large increase in melanoma among males was analysed in greater detail to investigate a suspected geographical link and was the focus of the remainder of the analysis. The advent of PSA testing combined with an aging population was thought to largely explain the spike in prostate cancer incidence while no obvious explanation for the melanoma figures could be found. As the primary risk factor for this cancer is behavioural (related to UV exposure) and largely avoidable; this cancer was chosen for further analysis and visualisation. Inspection of the melanoma data uncovered several unusual features that should be borne in mind. Firstly, several '0' values were reported for rate per 100,000; and where needed missing values were imputed by averaging the value for the preceeding & proceeding years ( Louth 2005, Leitrim 2005, Longford 2005). Secondly, while incidence was substantially higher at the end of study period, year on year rates varied drastically i.e. there was not a steady linear increase (mean 13.16 CPO, standard deviation = 5.07 CPO). This was presumably due to the low population density in many of these regions and limits our confidence in the reliability of county to county variation.

Selecting a colour & scale scheme for melanoma data proved extremely difficult. It was thought imperative to retain the same limits in colour scale for the three timepoints to allow direct comparison; but increasing the upper scale limit beyond 30 CPO resulted in the 1995 heatmap being completely uniform (all counties were considered to be at the lowest point of the scale). The increase was most apparent when the scale was shifted to a maximum of 30 CPO but this resulted in two counties (Kerry & Waterford) becoming greyed out as their incidence rates exceeded the upper limit of the scale. The out of bounds argument in dplyrs 'scale_fill_gradient' function was set to 'squish' which truncated these outlying values to the maximum of the scale. The colour scheme of yellow for low and red for high was intentionally chosen both because of the striking visual parallels with the sunburning process commonly accepted as a risk factor; and to emphasise the sheer severity of the increase in rate.

Mapping of the male melanoma incidence data highlights a marked nationwide increase in male melanoma with a pattern of particularly high incidence among coastal counties. This pattern appeared to support the intuitive hypothesis that proximity to beaches & water based recreational activities would result in higher levels of UV exposure. Given the size of Cork and how far it extends inland; it was also thought that the county-wide figures used to generate the image may have helped mask an effect in this region as the two neighbouring coastal counties Waterford and Kerry (51 CPO & 36 CPO) had the 1st & 2nd highest incidence respectively.

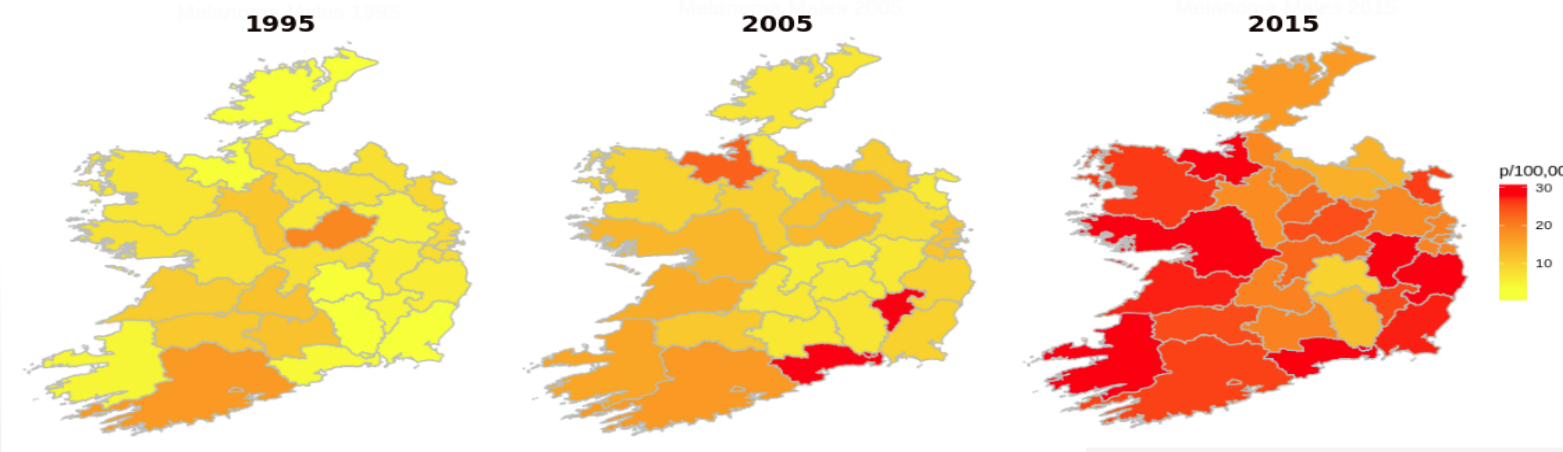



*Figure 5. Heatmap of male melanoma rate in each county.*

**Narrowing of the Gender Gap**

The final step in this analysis was to compare the rates of melanoma between the two sexes. Data misssing for rates in 1995 for Offaly (males), Sligo & Leitrim (females) were replaced with the national average as rates for the proceeding years were erratic and no robust method for imputing the values could be devised. Similarly; Longford 2015 was replaced with the average of the preceding year + 2015's national average. A potentially more robust method would have been to perform linear regression for incidence rates in Longford or multiply the 2014 rate by the average natonal increase between 2014 and 2015; but this was considered excessively time-consuming and the variance in yearly rates would also have limited confidence in this approach. The reasons underlying missing data were not apparent and appeared to be missing at random.

While statistical significance testing was not carried out; the uncovered effect was thought to be sufficiently large not to have been influenced by the handful of missing values (4 out of 52). Comparison of male & female melanoma rates revealed that while rates in women also increased notably during the study period (from 13.8 CPO in 1994 to 23.9 in 2015, or a 73.2% increase); the proportional increase in males was far more profoundover 300% and what was historically a female dominated cancer was now marginally more common in males. Data for each county were melted into a long format to allow comparison of rates both by year & by gender using base R's boxplot function for visualisation. Outliers above IQR*1.5 were removed to avoid squashing the 1994 data (potentially masking the disparity between male & female rates at the beginning of the study) and the plot fully captures the massive spike in male melanoma in Ireland since 1994 which was the most striking finding in this study (Figure 6).

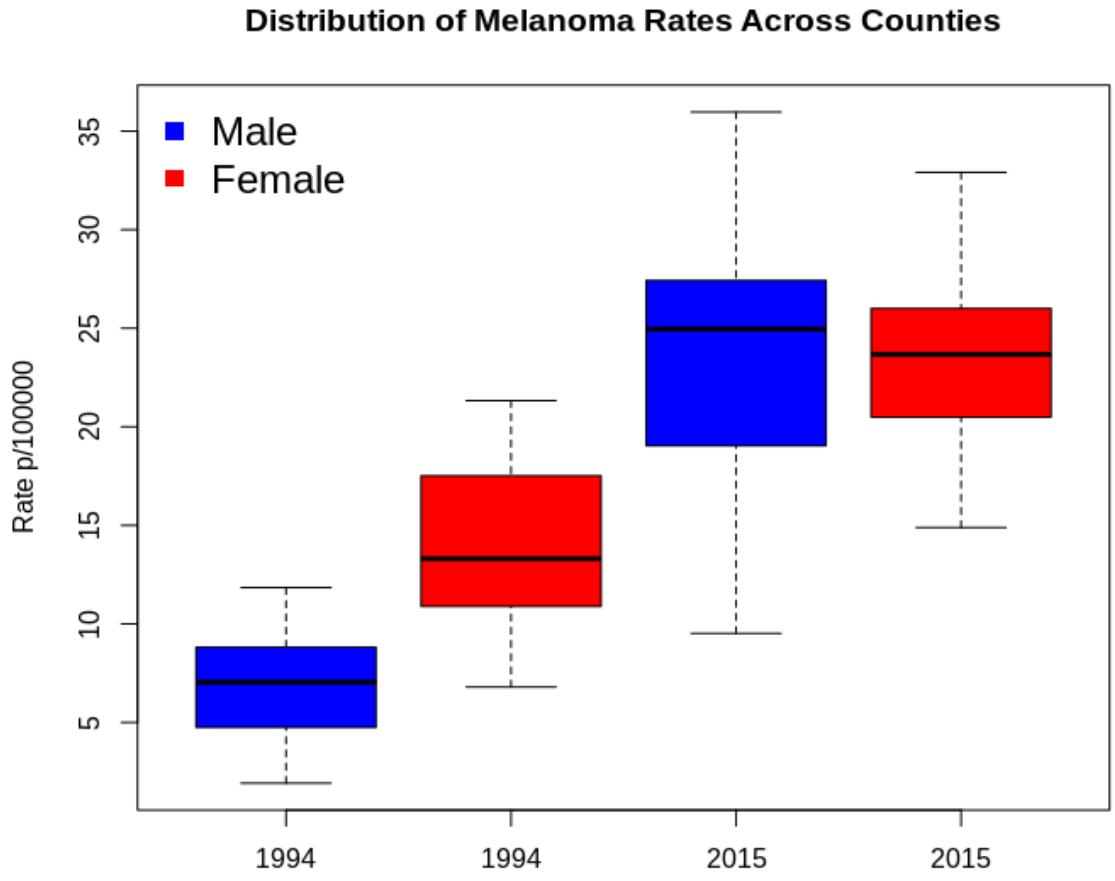



*Figure 6. Distribution of melanoma rates for 26 counties in 1994 & 2015 grouped by gender.*

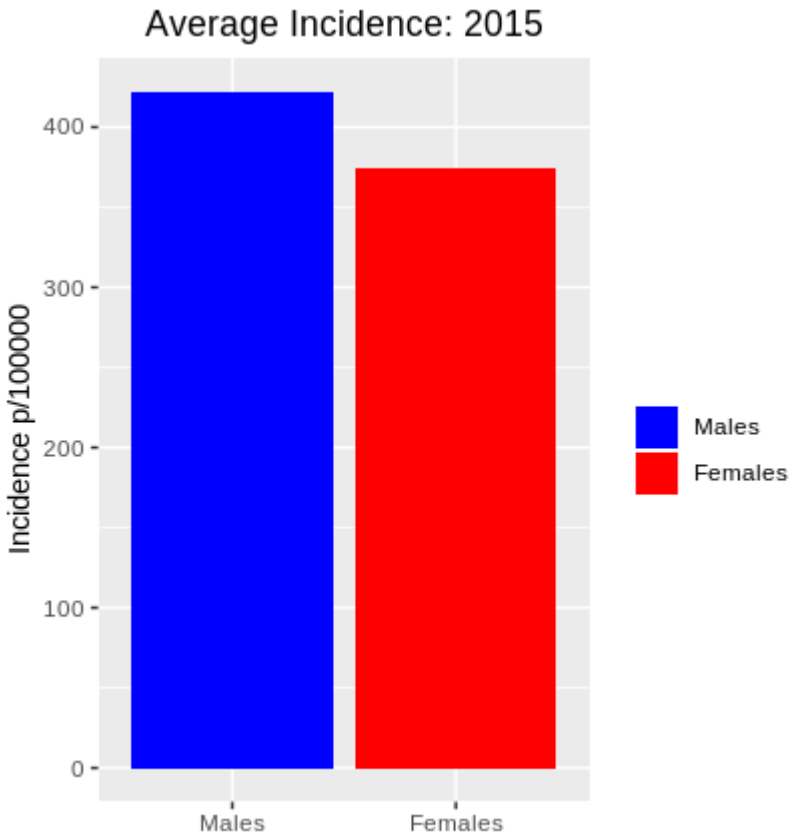
References:


1. https://www.ncri.ie/data/incidence-statistics.
2. https://www.nature.com/articles/1207717.
3. https://www.statista.com/statistics/376889/average-age-of-the-population-in-ireland/
4. http://rpubs.com/BrunoVoisin/csomaps

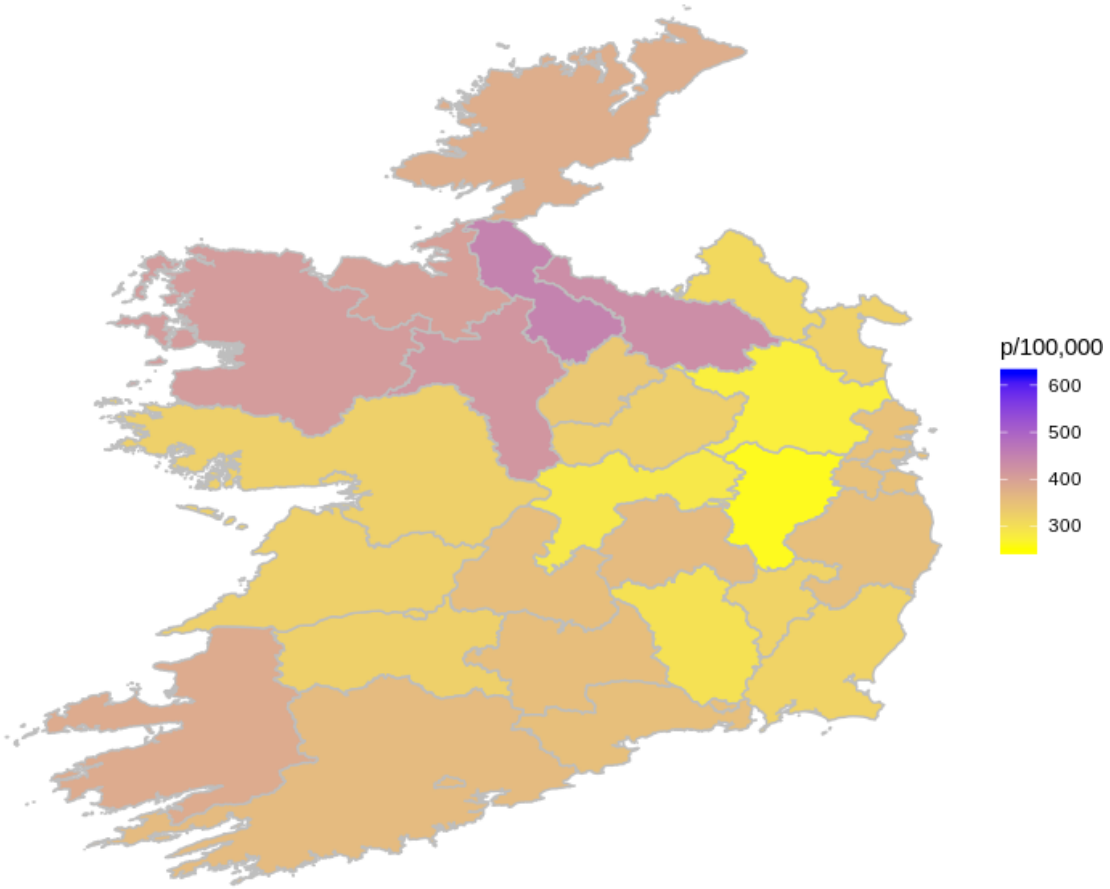

Unused Graphs and Tables

| | County | Rate.1994 | Rate.2015 |
|---|---|---|---|
| | <fct> | <dbl> | <dbl> |
| 1 | Leitrim | 449. | 551. |
| 2 | Cavan | 431. | 484. |
| 3 | Roscommon | 418. | 543. |
| 4 | Mayo | 409. | 600. |
| 5 | Sligo | 403. | 507. |
| 6 | Kerry | 384. | 534. |
| 7 | Donegal | 380. | 501. |
| 8 | Cork | 359. | 501. |
| 9 | Laois | 359. | 433. |
| 10 | Tipperary | 354. | 502. |

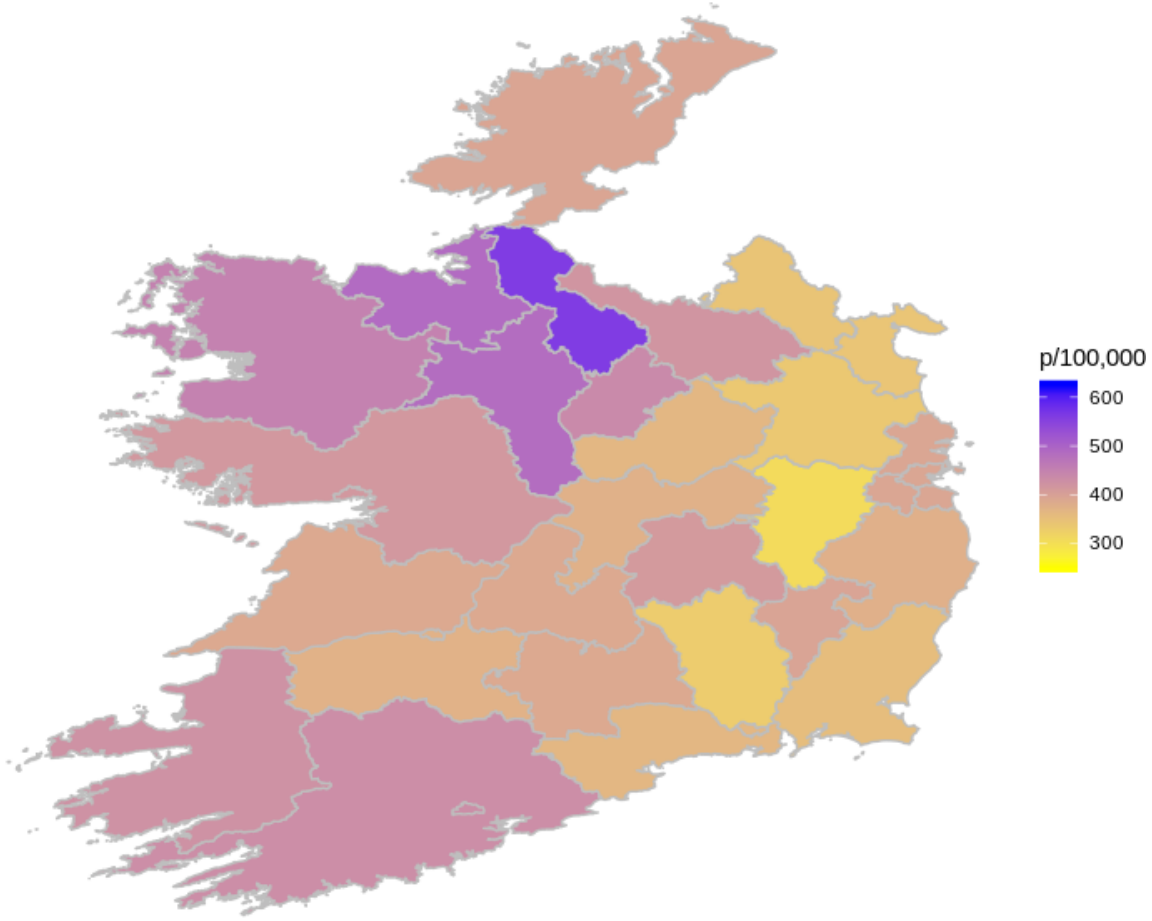
| | County | Rate.1994 | Rate.2015 |
|---|---|---|---|
| | <fct> | <dbl> | <dbl> |
| 1 | Mayo | 409. | 600. |
| 2 | Leitrim | 449. | 551. |
| 3 | Roscommon | 418. | 543. |
| 4 | Kerry | 384. | 534. |
| 5 | Louth | 323. | 521. |
| 6 | Sligo | 403. | 507. |
| 7 | Tipperary | 354. | 502. |
| 8 | Donegal | 380. | 501. |
| 9 | Cork | 359. | 501. |
| 10 | Cavan | 431. | 484. |



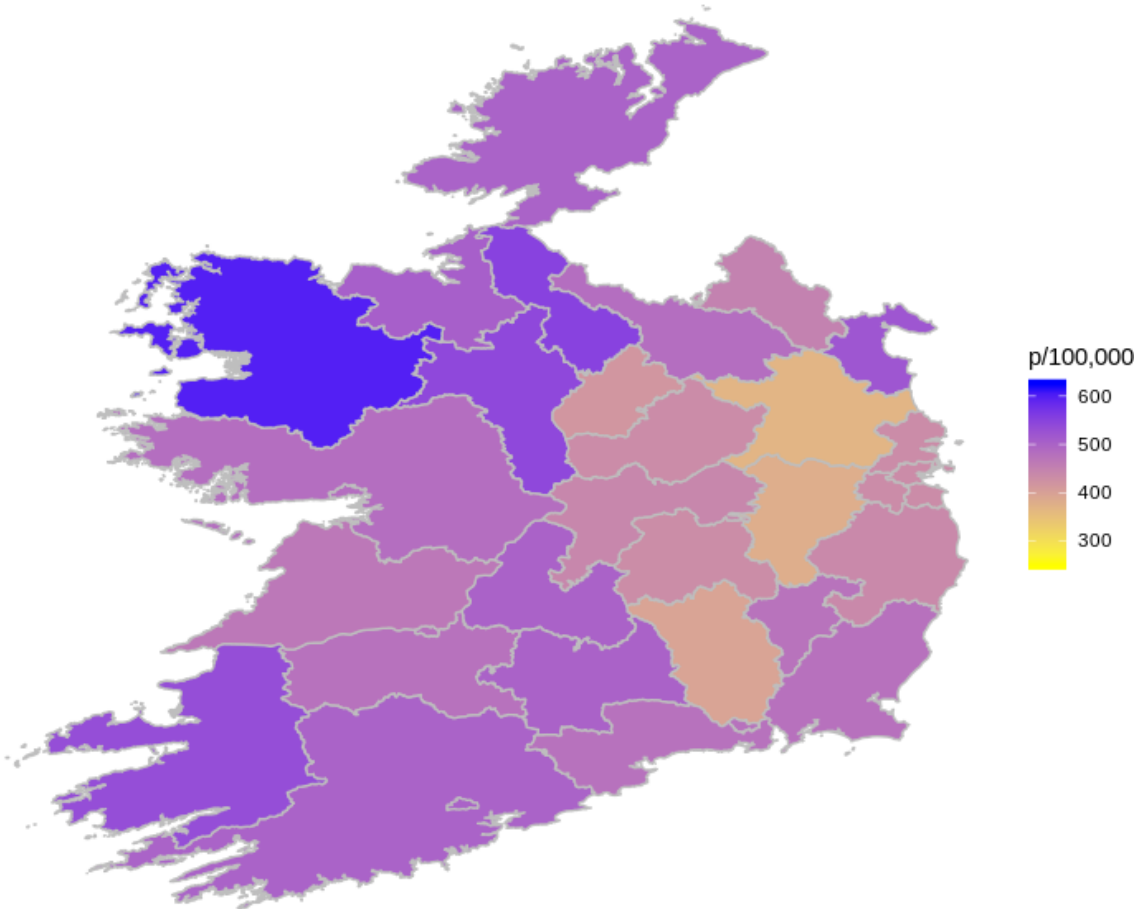
Overall Incidence 1995



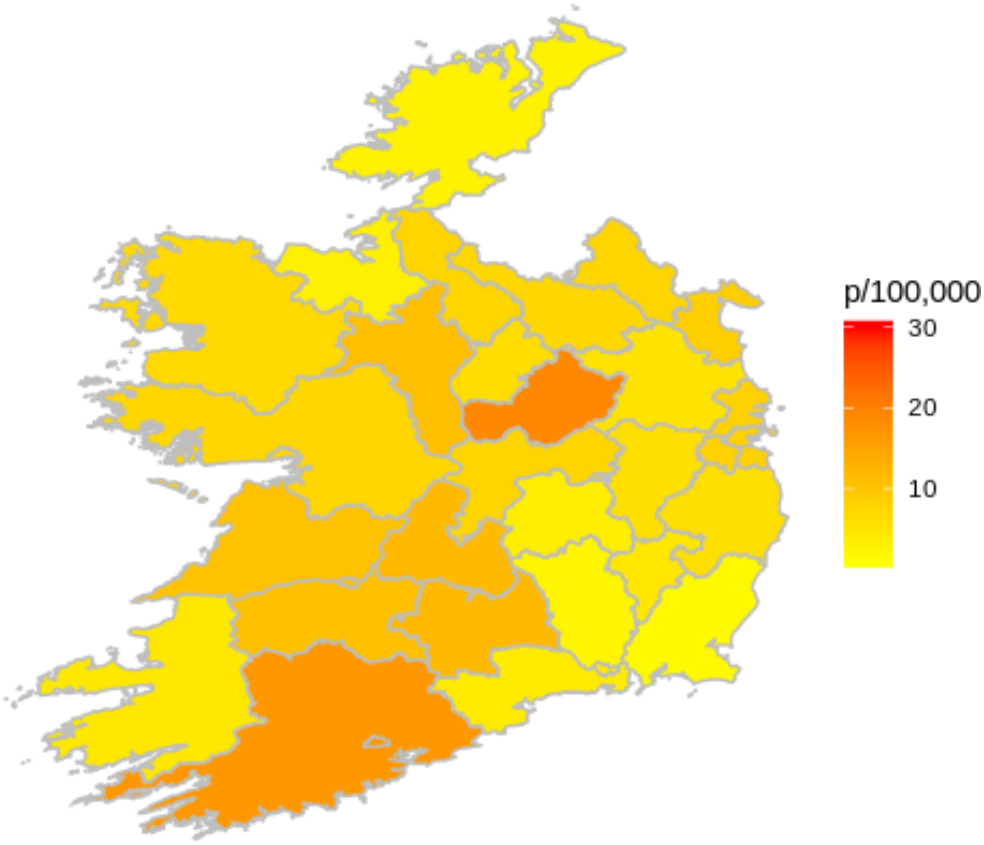

New Diagnoses 2005



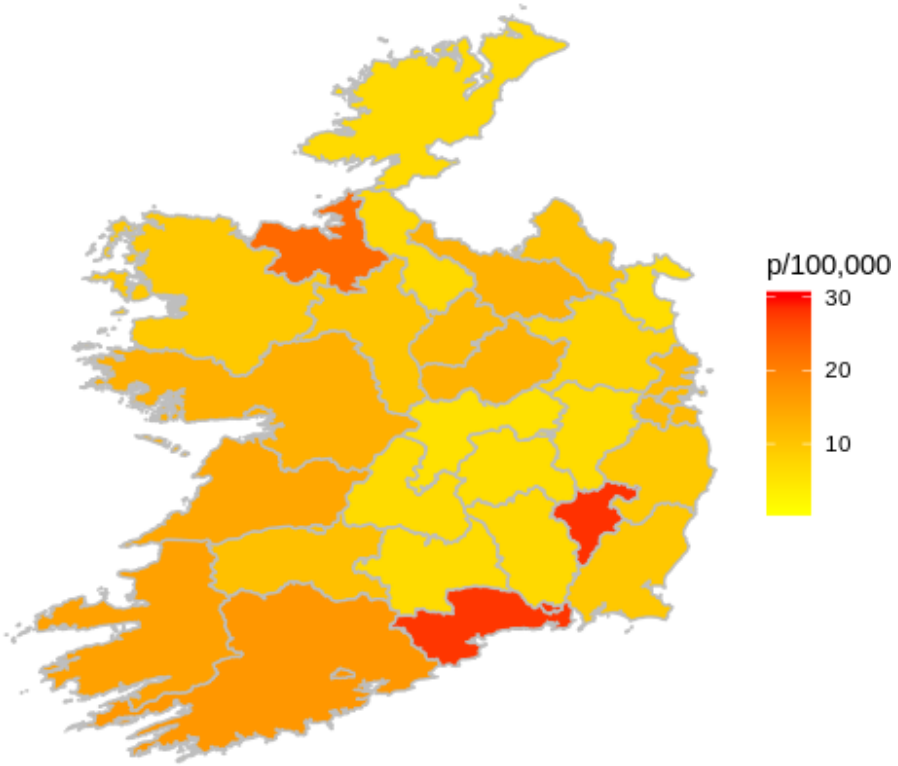
New Diagnoses 2015




Melanoma Males 1995



Melanoma Males 2005




Melanoma Males 2015