Cancer risk factors analysis in New Zealand

Todd (Zhen) Zhang, Shubo Feng, Dan Wei, Zichen Zhou

GitHub: https://github.com/Todd183/CancerRF

2023-10-16

1. Background

In the year 2020, New Zealand recorded a total of 27,072 new cancer registrations, reflecting an overall age-standardized rate of 338.1 registrations per 100,000 individuals.(Ora, 2023) This considerable volume of new cancer cases imposes a substantial socio-economic burden on the nation. However, it is worth noting that cancer is often far more manageable and treatable when detected at an early stage. This underscores the critical importance of cancer screening.

In New Zealand, a comprehensive national screening program is in place, covering breast, cervical, and bowel cancers, and it is accessible to all residents. For instance, the colorectal cancer (CRC) screening program is initiated at the age of 60 for the majority of the population. Nevertheless, an exception is made for Māori and Pasifika communities, who commence screening at 50 years of age.(Zealand, n.d.) This approach is founded on the fact that ethnicity has been established as a notable risk factor for CRC, with both Māori and Pasifika populations exhibiting a higher incidence of the disease.

The implementation of risk factors-adjusted cancer screening is essential for cost-effective cancer screening. Prior research has illuminated numerous social and environmental risk factors that are linked to cancer incidence. For instance, temperature has emerged as a significant environmental factor associated with the occurrence of cancer.(Voskarides, 2023) Also, studies have indicated that individuals with higher levels of education tend to have lower cancer rates.(Larsen et al., 2020)

This report collected and analyzed regional cancer incidence and risk factors, aiming at identifying potential risk factors associated with cancer incidence in New Zealand. By doing so, we aspire to provide valuable insights in refining and optimizing cancer screening strategies.

2. Methods

2.1 Data

To analyze cancer risk factors, two types of data were used in this report, including cancer and risk factors data. The cancer data includes incidence and mortality for different cancer types. The risk factors were classified into environmental and social factors. The environmental factors include earthquake, air quality, groundwater quality, and temperature, while the social factors contain income levels, working hours, highest educational qualification, and birth number. Birth number is a particular indicator on the female population and can be utilized for the analysis of women-related cancer.

Table 1: Data Sources

Type	Group	Category	Source
Cancer data	Incidence	Incidence	Te Whatu Ora
	Mortality	Mortality	
Risk Factors	Social factors	NZHS	Ministry of Health
		Education	Stats NZ
		Work Hours	
		Income	
		Birth numbers	
	Environmental factors	Earthquake	GeoNet
		Temperature	Stats NZ
		Air quality	LAWA
		Ground water quality	
Note. NZHS, New Zealand Health Survey; LAWA, Land, Air, Water Aotearoa.			

2.1.1 Cancer data

Cancer data is sourced from the Cancer Web Tool, which gathers the official data from the New Zealand Cancer Registry and New Zealand Mortality Collection. The original data is categorized based on DHB regions and documents cancer incidence for distinct types of cancer with gender information from 2011 to 2020. Due to the incidence rates of certain cancers are notably different between genders, we have grouped cancer types by gender based on the original information. This grouping is intended for subsequent analysis.

2.1.2 Risk factor data

2.1.2.1 Earthquake occurrences

The earthquake data was obtained from GeoNet, which offers detailed records of earthquakes that have occurred in New Zealand over the years. Given New Zealand's unique geographical location in a seismic zone, we investigate whether earthquake frequencies have an impact on cancer incidence. We have matched the latitude and longitude information in the original data to the corresponding DHB regions. Our primary focus lies on the annual highest and average values of earthquake magnitudes and depths, spanning from 2011 to 2020.

2.1.2.2 Air quality

Air quality data was collected from Land, Air, Water Aotearoa (LAWA). The original dataset provides latitude and longitude coordinates, monitoring site names, as well as concentrations of PM10 and PM2.5 from 2016 to 2022. PM10 particles have a diameter of less than 10 micrometers (µm), while PM2.5 particles are under 2.5 µm in diameter (LAWA, 2023a). We have converted the latitude and longitude coordinates into DHB regions and limited the time frame of our analysis to the years 2016 to 2020. Inhaling clean air is essential for our health (LAWA, 2023a). It is vital to consider air quality as a factor related to cancer incidence for analysis.

2.1.2.3 Groundwater quality

The data was also obtained from LAWA and encompasses groundwater quality monitoring in New Zealand from 2004 to 2021. Five indicators are utilized to assess groundwater quality, including the values of Chloride, Dissolved Reactive Phosphorus, Escherichia Coli, Electrical Conductivity, and Nitrate Nitrogen. The latitude and longitude information in the original dataset was converted into DHB regions, covering the period from 2011 to 2020. The groundwater is commonly utilized as an origin of drinking water (LAWA, 2023b). Therefore, we adopt this as an environmental risk factor related to cancer incidence.

2.1.2.4 Temperature

As one of the most perceptible climatic features to the human body, temperature has also been employed as one of the environmental risk factors. The dataset is derived from Statistics NZ. The original dataset recorded the highest and average temperatures in New Zealand, both seasonally and annually, from 1928 to 2022. We have selected data from the years 2011 to 2020 and converted the latitude and longitude information to corresponding DHB regions.

2.1.2.5 Working hours

The duration of working hours often brings varying levels of stress and significantly impacts personal health. Therefore, it has been employed as one of the risk factors in our cancer correlations analysis. The data is sourced from Statistics NZ, based on 2006, 2013, and 2018 Censuses. We selected year 2013 and 2018 and converted the area codes into DHB regions. Population proportions for different levels of working hours have been calculated for each DHB region, aiming to explore the relationship between working hours and cancer incidence.

2.1.2.6 Highest educational qualification

Educational level is potentially related to an individual's occupational background. For instance, a lower educational background may lead to a higher likelihood of engaging in physically demanding labour, which can impact physical health. The data is obtained from Statistics NZ based on the Censuses conducted in the years 2006, 2013, and 2018. We have transformed area codes into DHB regions and calculated the population proportions for each educational qualification level, with year 2013 and 2018. Cleaned dataset will be used for subsequent analysis to explore the influence of educational levels on cancer.

2.1.2.7 Income level

Income data is also sourced form Statistics NZ. Similarly, we have chosen data from the years 2013 and 2018, converted area codes into DHB region, and calculated the population proportions for each income level within each region. Income levels reflects the quality of life to some extent. We seek to investigate the relationship between income levels and cancer incidence.

2.1.2.8 Number of children born

The data is acquired from Statistics NZ, which presents the number of children women aged 15 and above have. We have computed the proportion of female population with different numbers of children for year 2013 and 2018. The cleaned data is utilized for conducting correlation analysis related to female-specific cancers, such as "Breast", "Ovarian", and "Uterine" cancer.

2.1.2.9 New Zealand health survey

This is a survey conducted by the Ministry of Health from 2011 to 2019, focusing on individuals' health habits and health status. We utilized the relevant indicators from this survey to conduct a correlation analysis of cancer incidences. These indicators include smoking and drinking habits, dental health, BMI, and health

insurance, etc. A limitation of this data is that in the original data, some variables only provide descriptions of the degree, such as "heavy smokers", which are based on scores obtained from the survey. This means that we do not have specific value to describe how many cigarettes are smoked per day would be categorized as "heavy smokers". However, it is worth noting that the adjective can effectively convey the degree of health habits.

Data sources are shown in **Table 1**. Detailed descriptions of each dataset and variables are documented in: "*5. Data_Documentation".*

2.2 Analysis Pipline

2.2.1 Data Pre-processing

2.2.1.1 Cancer Datasets Processing

The primary objective of data processing is to transform the dataset into a format with "DHB" and "Year" as the common identifiers for ease of subsequent data joins. To achieve this, a systematic approach was applied to the datasets, encompassing cancer incidence and mortality data. The DHB text was standardized using regular expressions to ensure consistency. Simultaneously, we handled missing values by filling them with 0. This is primarily because the data pertains to incidence rates, and the low number of cases for certain diseases renders these missing values statistically insignificant.

After that, we narrowed down our focus to the specific types of cancer we were interested in by choosing the rates at which they occur based on whether the person is male or female. This helps us analyze cancer incidence in a way that's directly relevant to what we're studying. The data is now organized by "Year" and "DHB," making it ready for easy combination and detailed analysis.

2.2.1.2 Risk Factors Datasets Processing

In the Risk Factors datasets, we handled geographical data, addressed missing and duplicate entries, and extracted features through data grouping. Specifically, for datasets containing geographical information, a key step involved converting latitude and longitude data into their corresponding DHB regions.

Additionally, addressing instances of multiple entries in a single year, like recurring surveys, we opted for averaging to streamline the information. Moreover, comprehensive processing of weather data involved extracting quarterly averages, maximum, and minimum values. Similarly, factors such as education levels and income were synthesized by aggregating data, providing a more consolidated and useful representation of these variables.

Details of data wrangling process and codes are recorded in Jupyter notebooks:

- *1.1. Data_wrangling_Cancer_Data_Julila.ipynb*
- *1.2. Data_wrangling_Risk_Factors_R.ipynb*

2.2.2 Cancer Overview

The first part of our analysis centers on assessing the broader cancer landscape in New Zealand over the last decade. We adopt a multi-dimensional approach, examining the data from five distinct angles: age distribution, gender disparities, regional patterns, and temporal trends. Within the age distribution analysis, we present cancer incidence rates for both pediatric and adult groups, aligning with the age threshold used in the NZHS data—dividing individuals into "children" (< 14 years old) and "adults" (14 years old). Our exploration of gender variance seeks to quantify the distinctions in both cancer incidence and mortality between male and female patients across various cancer types.

2.2.3 Correlation analysis

Correlation analysis between cancer incidence and risk factors is performed on regional (DHB) level. We focus on conducting linear regression analysis on cancer incidence rates in various regions of New Zealand over the past decade. We treat different DHBs and years as individual observations. Consequently, we have approximately 40-200 data points for each type of cancer, with a total of 20 DHBs. When we have complete data for 10 years, we obtain 20*10 observations. Our datasets contain a minimum of 2 years of data.

In our correlation analysis, we use the Pearson correlation coefficient. While we have chosen multiple features for each type of characteristic, they are not mutually independent. For example, we consider both annual and seasonal average temperatures, as well as the highest and lowest temperatures. We have applied the Benjamini-Hochberg method to adjust p-values for multiple statistical tests. Additionally, we retained the total number of features within each category. We assume that factors from different category can be considered independent variables, so when adjusting p-values, we use the number features within the same category.

For visualization purposes, we use a volcano plot to represent the data. Significance is indicated by the negative logarithm of the p-value, and the correlation degree is directly represented by the Pearson correlation coefficient.

3. Results

3.1 Cancer Overview

3.1.1 Age distribution

Table 2: Number of New Registered Cancers

population	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
adult	63,678	66,002	67,046	69,662	70,190	72,881	73,990	78,844	79,357	80,944
child	252	310	310	256	298	304	248	272	284	272

The number of adult cancer registrations is over 800 to 1000 times greater than the number of child cancer registrations.

This table presents cancer registration data from 2011 to 2020, categorized into adults and children. While there has been an upward trend in adult cancer registrations, child cancer registrations have remained relatively stable.

Given the low proportion of child cancer registrations in the overall dataset, the small percentage of child data does not significantly impact our analysis. And our analysis primarily focuses on the general cancer situation in New Zealand. Furthermore, the cancer data we've collected from various regions doesn't differentiate between adults and children. As a result, we can't analyze adult data separately in subsequent analyses.

3.1.2 Gender variance

Females had the highest average incidence of "Breast cancer" from 2011 to 2020, while males had the highest for "Prostate cancer". "Colorectal cancer" also showed high average incidence rates for both genders.

For average mortality rates over the same period, both genders exhibited the highest rates from "Lung cancer". For females, "Breast cancer" was the second-leading cause of mortality, while for males, it was "Colorectal cancer".

Figure 1 represents these average incidence and mortality rates for each cancer type by gender based on data from the website https://tewhatuora.shinyapps.io/cancer-web-tool/. However, some data points were missing: female data for "Kidney cancer" and "Bladder cancer" and male data for "Thyroid cancer".

Cancers exclusive to one gender, like "Uterine cancer", "Ovarian cancer", and "Breast cancer" in females and "Prostate cancer" and "Testicular cancer" in males, were analyzed separately. For other cancer types with data for both genders, separate analyses were conducted. Generally, average incidence and mortality rates were similar between genders, though males had slightly higher incidence rates.

Among cancer types with data for both sexes, the average incidence and mortality rates were generally similar between males and females, with slightly higher average incidence rates among males. In subsequent analyses, cancers with data for both genders were analyzed using "all sex" data, while cancers with data available for only one gender were analyzed using the corresponding gender-specific data.

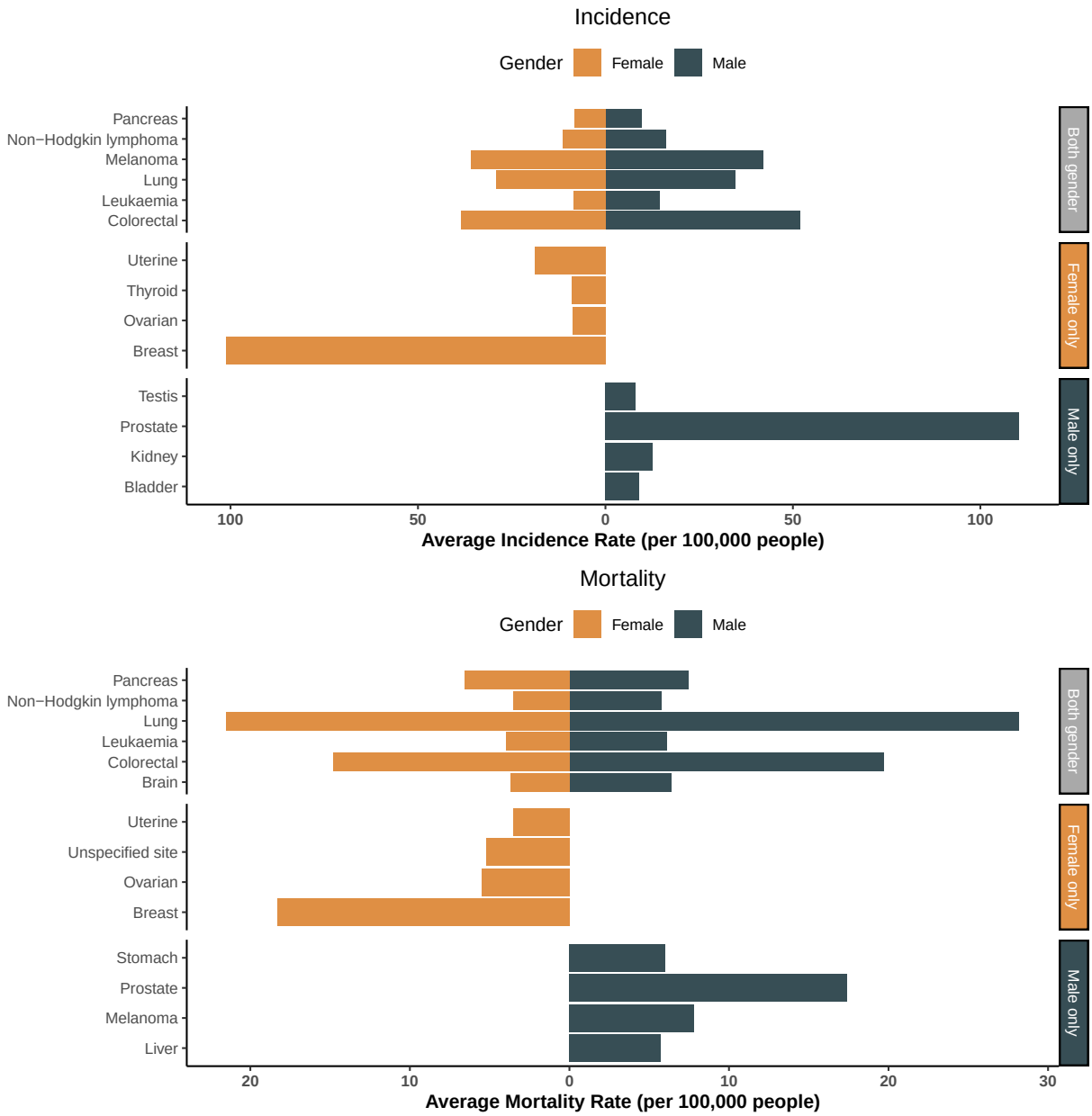

Figure 1: Average Incidence Rate and Average Mortality Rate for each cancer type by sex

3.1.3 Regional distribution

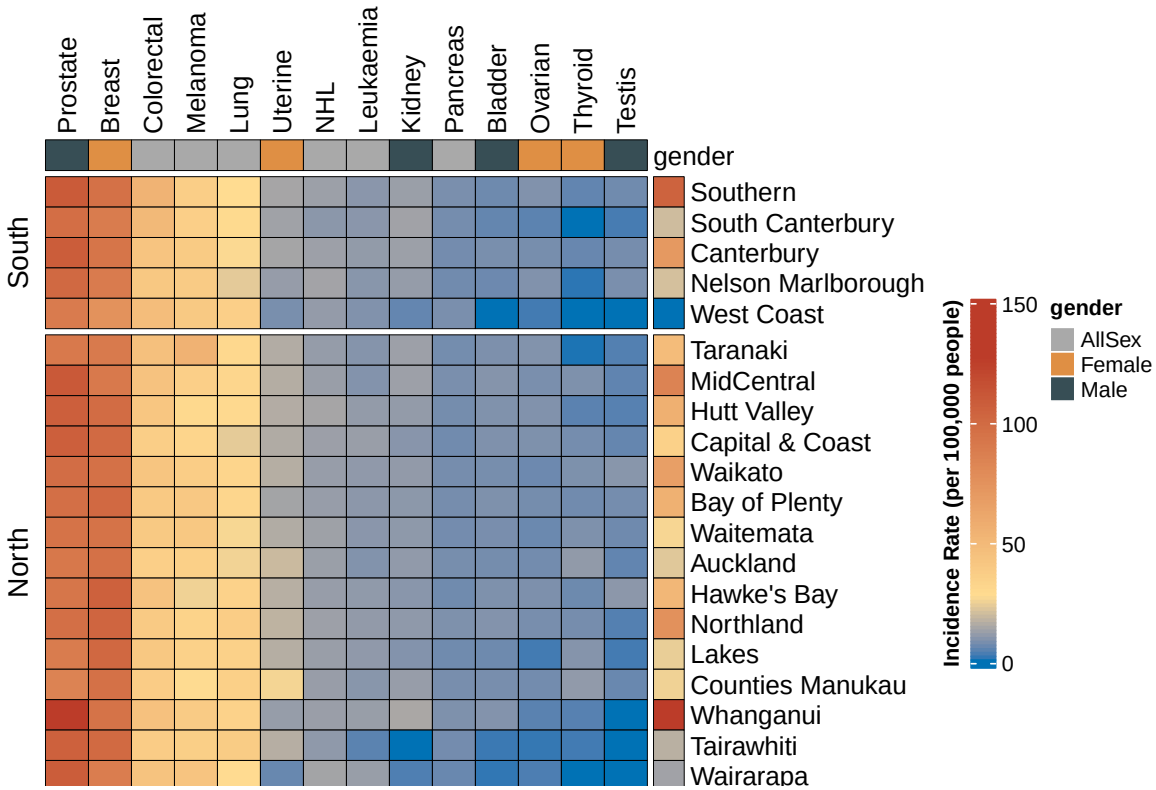

Figure 2: Heatmap illustrating the regional distribution of average cancer incidence during 2011 to 2020 in New Zealand. Gradient color indicated the average cancer incidence rate for the corresponding region.

From 2011 to 2020, Whanganui had the highest cancer incidence rate among all New Zealand regions, with the West Coast recording the lowest.

Figure 2 uses shades of orange to represent higher cancer incidence rates and shades of blue for lower rates. Whanganui and Northland had notably high rates, whereas the West Coast had the lowest. It also shows that "Prostate cancer" and "Breast cancer" had the highest incidence rates in every region, while "Thyroid cancer" and "Testicular cancer" had the lowest rate.

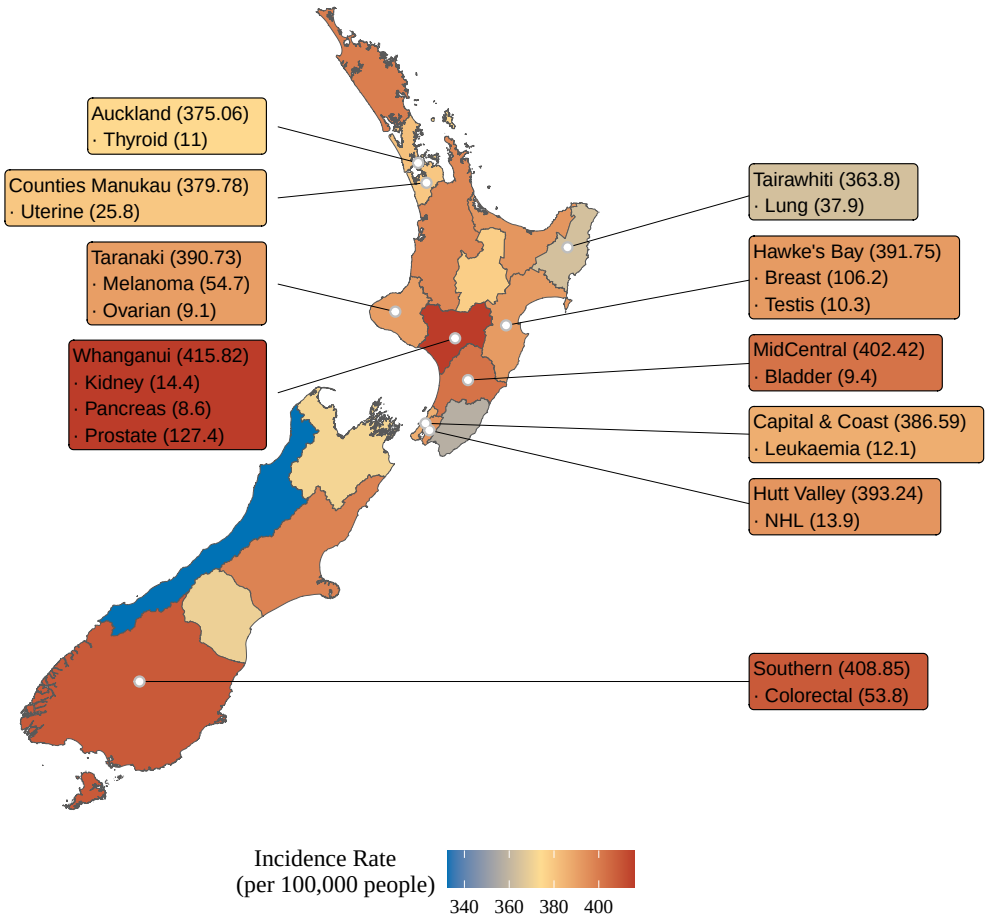

Figure 3: Regional distribution of overall cancer incidence in New Zealand. Gradient color indicated the average cancer incidence rate (All cancer types) for the corresponding region during 2011 to 2020. Lables highlight the most common region for corresponding cancer type

Figure 3 provides a clearer depiction of regional cancer rates. Whanganui topped the list with an incidence rate of 415.82/100,000, primarily driven by "Kidney cancer", "Pancreas cancer", and "Prostate cancer". The Southern region followed with 408.55/100,000, where "Colorectal cancer" was most prevalent. The West Coast had the lowest incidence rate.

3.1.4 Tempory trends

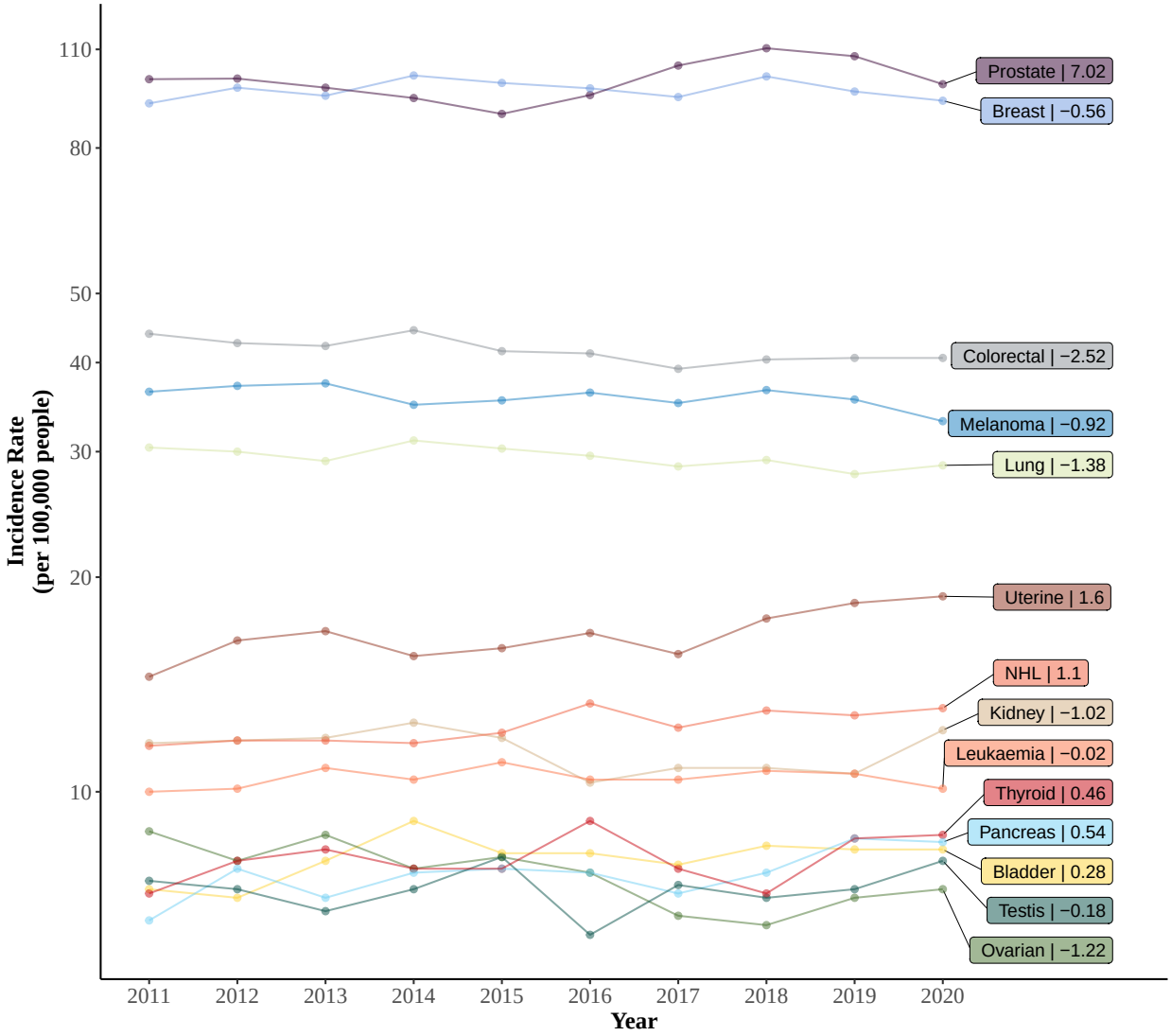

Figure 4: Temporal trends for cancer incidence of each cancer type during 2011 to 2020. The number in each label indicate the difference between average incidence of 2016-2020 and 2011-2015.

Between 2011 and 2022 year, the incidence rates of "Prostate cancer", "Uterine cancer", "Non-Hodgkin lymphoma cancer", "Thyroid cancer", "Pancreas cancer", and "Bladder cancers" all increased. In contrast, "Colorectal cancer" showed a significant decline. Lastly, the incidence rates for "Thyroid cancer" and "Testicular cancers" fluctuated substantially during this decade, while other cancers had more stable trends.

3.2 Correlation analysis

3.2.1 Correlation Overview

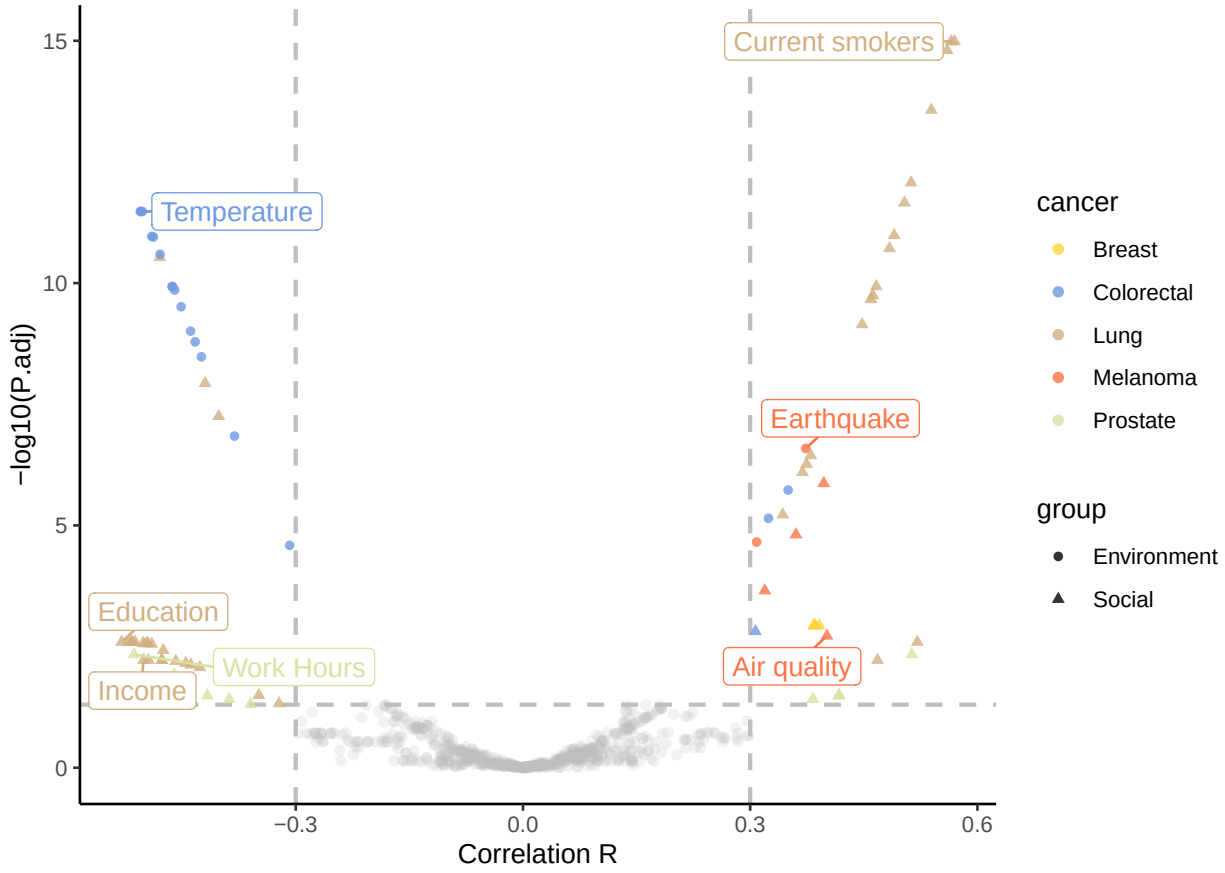

Figure 5: Risk Factor Revelations: Correlation and Significance of Risk Factors Across Five Types of Cancers

For the purpose of presentation and analysis, we have chosen to showcase the top 5 cancer types based on their incidence rates. **Figure 5** illustrates the correlation and significance of various risk factors with these cancers.

The y-axis represents the logarithmically transformed p-value, where higher values indicate greater significance, while the x-axis corresponds to the Pearson correlation coefficient. The plot distinctly highlights numerous risk factors strongly associated with specific cancers. For instance, there is a significant and positive correlation between current smokers and "Lung Cancer," aligning with our expectations. This serves as evidence that our sampling and analysis methods are, to a certain extent, reasonable.

The plot reveals intriguing findings, such as a positive correlation between earthquakes and air quality with "Melanoma Cancer". Additionally, there is a significant negative correlation between temperature and "Colorectal". These findings not only enrich our understanding of specific risk factor relationships but also present opportunities for timely interventions in cancer screening and prevention.

3.2.2 Risk Factors of Cancers

Table 3: Risk Factors of Breat, Prostate, Colorectal Cancer

Cancer	Correlation	Group	Category	Risk Factors	*r*	p.adj
Breast	+	Social	NZHS	Heavy Episodic Drinking at Least Weekly (total population)	.39	0.00
Prostate	+	Social	Income	> $10,000	.42	0.03
Prostate	-	Social	Work Hours	≥ 10 Hours	-.51	0.00
Colorectal	+	Social	Air quality	PM2.5 Concentration Max	.34	0.01
Colorectal	+	Environment	Earthquake	Magnitude Mean	.35	0.00
Colorectal	+	Social	NZHS	Past-Year Drinkers	.31	0.00
Colorectal	-	Environment	Temperature	Average Annual	-.50	0.00

Table 3 describes the risk factors associated with "Breast Cancer", "Prostate Cancer", and "Colorectal Cancer".

Notably, a significant positive correlation is observed between "Breast Cancer" and Heavy Episodic Drinking, emphasizing the potential influence of alcohol consumption on breast cancer risk. Conversely, "Prostate Cancer" demonstrates a strong negative correlation with extended work hours (weekly work hours 10), suggesting a potential protective effect associated with shorter work durations.

"Colorectal Cancer" shows positive correlations with three factors: maximum PM2.5 concentration, average earthquake magnitude, and being a "Past-year drinker". These findings illuminate the diverse factors influencing colorectal cancer risk, including environmental pollutants, seismic activity, and past-year drinking behavior.

Table 4: Risk Factors of Lung cancer

Cancer	Correlation	Group	Category	Risk Factors	*r*	p.adj
Lung	+	Social	NZHS	Current Smokers	.57	0.00
Lung	+	Social	NZHS	High Blood Pressure (medicated)	.34	0.00
Lung	+	Social	NZHS	Little or No Physical Activity	.46	0.00
Lung	+	Social	NZHS	Mean BMI (kg/m²)	.50	0.00
Lung	+	Social	NZHS	Only Visit Dental Health Care Worker for Problems	.57	0.00
Lung	-	Social	Education	≥ Level 4	-.53	0.00
Lung	-	Social	Income	> $30,000	-.50	0.01
Lung	-	Social	NZHS	Dental Health Care Worker Visit	-.40	0.00
Lung	-	Social	NZHS	Private Health Insurance	-.48	0.00

Table 4 shows factors associated with "Lung Cancer". Firstly, it is evident that being a Current Smoker significantly increases the risk of lung cancer. Additionally, weight-related factors such as being Obese and having a higher Mean BMI (kg/m²) are correlated with an elevated incidence of lung cancer. Similarly, engaging in Little or No Physical Activity and having High Blood Pressure (medicated) are positively associated with lung cancer.

Three factors related to Dental Health show correlations with lung cancer. "Dental Health Care Worker Visit" exhibits a negative correlation, indicating a potential protective effect, while "Only Visit Dental Health Care

Worker for Problems" and "Teeth Removed Due to Decay in Lifetime" are positively correlated with lung cancer. Furthermore, income-related factors are linked to lung cancer, with higher income, higher education levels, and having Private Health Insurance showing negative correlations with lung cancer incidence.

Table 5: Risk Factors of Melanoma

Cancer	Correlation	Group	Category	Risk Factors	*r*	p.adj
Melanoma	+	Social	Air quality	PM2.5 Concentration Mean	.40	0.00
Melanoma	+	Environment	Earthquake	Depth Mean	.37	0.00
Melanoma	+	Social	NZHS	All Teeth Removed Due to Decay	.36	0.00
Melanoma	+	Social	NZHS	Mean Height (cm)	.40	0.00
Melanoma	+	Social	NZHS	Self-Rated Health - Very Good	.32	0.00

Table 5 shows the risk factors linked to Melanoma. Within environmental considerations, Melanoma exhibits a positive correlation with the average concentration of PM2.5 and the average depth of earthquakes. In the realm of social factors, the incidence of Melanoma is positively linked to "All Teeth Removed Due to Decay". Furthermore, factors like "Self-Rated Health - Very Good" and "Mean Height" also display positive correlations with Melanoma incidence.

3.2.3 Distribution in Detailed Observations

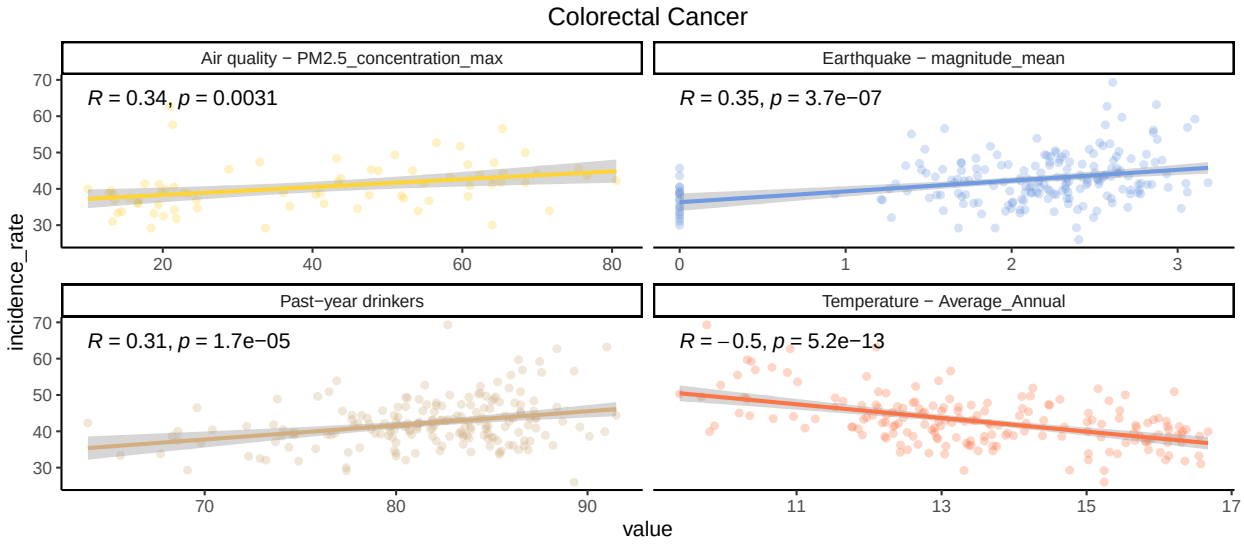

Figure 6: Correlation between colorectal cancer and significant risk factors

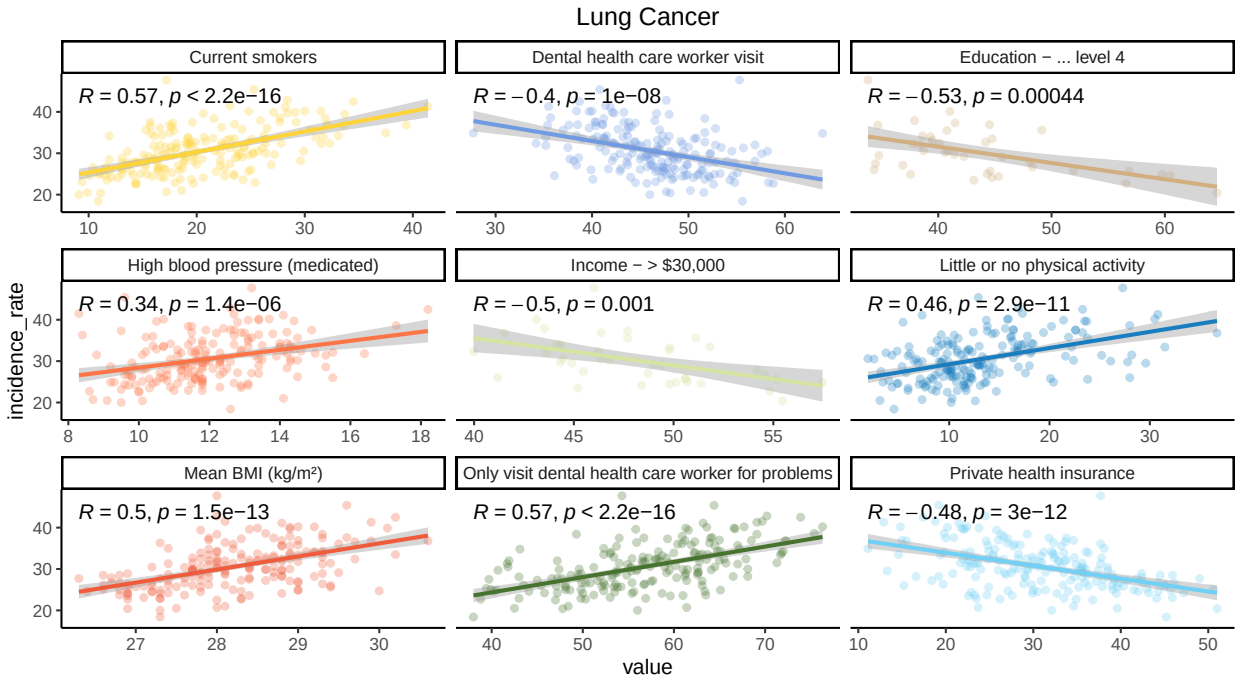

Figure 7: Correlation between lung cancer and significant risk factors

Figures 6 and **Figure 7** illustrate the relationship between incidence rates and key risk factors for "Colorectal" and "Lung Cancer", providing a detailed analysis.

Notably, for both cancers, the graphs reveal discernible correlations with risk factors. While the seismic factor appears more scattered on the left side due to regions with no seismic activity, other distributions demonstrate pronounced correlations, with concentrated patterns across the graphs. Significantly, the distribution of Lung Cancer in relation to weight exhibits a conspicuous linear pattern, indicating a noteworthy correlation between these variables.

4. Discussion

Our analysis delves into the multifaceted landscape of cancer risk factors, including various cancer types in New Zealand. The results we present not only corroborate prior research findings but also introduce new dimensions to our understanding of cancer etiology, underscoring the importance of considering environmental and social factors in cancer prevention strategies.

Colorectal Cancer Risk Factors: Our findings affirm well-established associations between PM2.5 (Ku et al., 2021), cold exposure(Voskarides, 2023), or alcohol drinking (Rossi, Anwar, Usman, Keshavarzian, & Bishehsari, 2018) and increased colorectal cancer risks. This connection emphasizes the pivotal role of environmental factors, specifically air quality, in shaping colorectal cancer risk. Furthermore, the inclusion earthquakes as a risk factor for colorectal cancer, merits further investigation, especially in earthquake-prone regions like New Zealand. The association between earthquakes and colorectal cancer may be attributed to environmental changes that occur following seismic activity.

Lung Cancer Risk Factors: Our analysis reveals risk factors for lung cancer, including overweight (Vedire, Kalvapudi, & Yendamuri, 2023), lack of physical activity (Cannioto et al., 2018), hypertension (Lindgren, 2003), and smoking (Walser et al., 2008), which align with well-documented risk factors for this cancer type. Lifestyle choices and health status are crucial contributors to lung cancer risk. Moreover, the inclusion of low income and education levels underscores the socio-economic determinants of lung cancer, as these factors

are closely associated with smoking habits (Larsen et al., 2020). The surprising link between dental health care visits and lung cancer emphasizes the significance of oral health in cancer prevention. Several studies have demonstrated that poor oral health is linked to an increased risk of lung cancer (Yoon et al., 2019), suggesting the importance of regular dental check-ups for maintaining oral health and potentially reducing lung cancer risk.

Melanoma Risk Factors: For melanoma, our analysis highlights risk factors such as PM2.5 concentration and average annual earthquake depth, indicative of the role of environmental factors in this skin cancer. Additionally, teeth loss and mean height are intriguing findings that merit further exploration. Understanding the mechanisms behind these associations is essential to develop effective strategies for melanoma prevention in New Zealand.

Breast Cancer and Drinking: The inclusion of heavy episodic drinking as a risk factor for breast cancer, as indicated by previous studies such as (McDonald, Goyal, & Terry, 2013), underscores the complexity of breast cancer etiology. This finding emphasizes the importance of addressing lifestyle factors, including alcohol consumption, in breast cancer prevention efforts.

Prostate Cancer and Work Hours: While our analysis suggests that lower work hours may be a risk factor for prostate cancer, it's important to acknowledge a potential bias. Lower work hours could be associated with older age, a known vulnerability factor for prostate cancer. However, due to the absence of age data in regional cancer incidence, further age-specific analysis is needed to confirm or rule out this bias.

Our analysis has limitations, including the inability to conduct subgroup analyses based on ethnicity or age group due to the lack of relevant data, potentially hindering our understanding of demographic-specific cancer incidence variations. Additionally, our correlation analyses were executed at the regional level, whereas comparing risk factors between cancer and non-cancer individuals at the individual level would have yielded more accurate results, although such data were not publicly available. Furthermore, some data sources lacked comprehensive variable documentation, introducing ambiguity in our interpretation of certain risk factors, such as "heavy episodic drinking" from the New Zealand Health Survey, for which a precise definition was absent.

In summary, our findings expand our knowledge of cancer risk factors for various types of cancer in New Zealand. These results underscore the multifaceted nature of cancer etiology and the importance of considering environmental and social factors in cancer prevention strategies. Our findings can serve as a foundation for the development of targeted prevention and screening strategies tailored to the unique risk factors associated with different cancer types, ultimately contributing to a reduction in the cancer burden in New Zealand. Further research on less-explored risk factors, such as earthquake-related cancer risks, will be crucial in advancing our understanding and improving prevention efforts.

5. References

Cannioto, R., Etter, J. L., LaMonte, M. J., Ray, A. D., Joseph, J. M., Qassim, E. A., … Moysich, K. B. (2018). Lifetime physical inactivity is associated with lung cancer risk and mortality. *Cancer Treatment and Research Communications*, *14*, 37–45. https://doi.org/10.1016/j.ctarc.2018.01.001

Ku, M.-S., Liu, C.-Y., Hsu, C.-Y., Chiu, H.-M., Chen, H.-H., & Chan, C.-C. (2021). Association of ambient fine particulate matter (PM _{2.5}) with elevated fecal hemoglobin concentration and colorectal carcinogenesis: A population-based retrospective cohort study. *Cancer Control*, *28*, 107327482110412. https://doi.org/10.1177/10732748211041232

Larsen, I. K., Myklebust, T. Å., Babigumira, R., Vinberg, E., Møller, B., & Ursin, G. (2020). Education, income and risk of cancer: Results from a norwegian registry-based study. *Acta Oncologica*, *59*, 1300–1307. https://doi.org/10.1080/0284186X.2020.1817548

LAWA. (2023a). *Factsheet: Why is air quality important?* Retrieved from https://www.lawa.org.nz/learn/factsheets/air-quality-topic/why-is-air-quality-important/

LAWA. (2023b). *Groundwater quality.* Retrieved from https://www.lawa.org.nz/explore-data/groundwater-quality/

Lindgren, A. (2003). Blood pressure, smoking, and the incidence of lung cancer in hypertensive men in north karelia, finland. *American Journal of Epidemiology*, *158*, 442–447. https://doi.org/10.1093/aje/kwg179

McDonald, J. A., Goyal, A., & Terry, M. B. (2013). Alcohol intake and breast cancer risk: Weighing the overall evidence. *Current Breast Cancer Reports*, *5*, 208–221. https://doi.org/10.1007/s12609-013-0114-z

Ora, T. W. (2023). *Cancer web tool*. Retrieved from https://www.tewhatuora.govt.nz/our-health-system/data-and-statistics/nz-health-statistics/health-statistics-and-data-sets/cancer-data-and-statistics/cancer-web-tool

Rossi, M., Anwar, M. J., Usman, A., Keshavarzian, A., & Bishehsari, F. (2018). Colorectal cancer and alcohol consumption—populations to molecules. *Cancers*, *10*, 38. https://doi.org/10.3390/cancers10020038

Vedire, Y., Kalvapudi, S., & Yendamuri, S. (2023). Obesity and lung cancer—a narrative review. *Journal of Thoracic Disease*, *15*, 2806–2823. https://doi.org/10.21037/jtd-22-1835

Voskarides, K. (2023). The double face of cold in cancer. *Translational Oncology*, *28*, 101606. https://doi.org/10.1016/j.tranon.2022.101606

Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., … Dubinett, S. M. (2008). Smoking and lung cancer: The role of inflammation. *Proceedings of the American Thoracic Society*, *5*, 811–815. https://doi.org/10.1513/pats.200809-100TH

Yoon, H.-S., Wen, W., Long, J., Zheng, W., Blot, W. J., & Cai, Q. (2019). Association of oral health with lung cancer risk in a low-income population of african americans and european americans in the southeastern united states. *Lung Cancer*, *127*, 90–95. https://doi.org/10.1016/j.lungcan.2018.11.028

Zealand, B. C. N. (n.d.). *New zealand's national bowel screening*. Retrieved from https://bowelcancernz.org.nz/about-bowel-cancer/early-detection-and-prevention/screening/

Answer to questions

Q1: What data source you used?

Cancer incidence data: This data presents cancer incidence with different cancer types and gender information, and was obtained from Cancer Web Tool, which presents the official cancer data from the New Zealand Cancer Registry and New Zealand Mortality Collection.

Cancer risk factors data: This data includes environmental factors and social factors. Environmental factors contain earthquake occurrences, air quality, groundwater quality, and temperature, while social factors encompass income level, educational qualification level, working hours, and the number of children born, as well as survey data on individual health status. These data originate from various sources. The Earth quake data was sourced from GeoNet, which offers earthquake details in New Zealand. The air quality and groundwater quality data were acquired from Land, Air, Water Aotearoa (LAWA), while the data of temperature, working hours, educational level, and number of children born were all sourced from Statistic NZ. The NZ health survey was obtained from Ministry of Health.

Q2: Why you chose those data sources?

Authenticity and reliability: All data comes from official and authoritative sources to ensure the authenticity and reliability of our study.

Valuable for our study: On the one hand, the data of cancer incidences can be used to analyze main trends of different types of cancer. On the other hand, we focused on two key directions of risk factors: environmental and social factors. Regarding environmental factors, we have selected representative factors such as earthquake occurrences, air quality, groundwater quality, and temperature. These factors are all crucial to our environment and can impact human health. As for social factors, the factors that exhibit potential influences on human health are taken into account. More specifically, the number of children born, as a unique indicator specific to women, can be employed to analyze cancers that primarily affect females. The working hour and highest qualification reflect the occupational background, which is closely related to

individual's health, while income reflects the economic status and their quality of life, which may have an impact on health.

Q3: What target you chose?

Examine the trends in the incidence of various cancer types: By utilizing our data and conducting the overview analysis, we examined the differences in cancer incidences among adults and children, as well as between males and females, and distinct features of cancers across different regions.

Cancer correlation analysis: We intended to subsequently conduct correlation analysis to identify the environmental and social factors that exhibit a strong associations with the incidence of cancer in each region over the years.

Q4: What difficulties you had to overcome to wrangle the data sources into the target data model?

Mapping Coordinates to DHB Regions: Our research focused on the relationship between cancer occurrences and socio-environmental factors, utilizing "region and year" as a common identifier. While cancer data was available in "DHB regions," other datasets primarily used "council regions." Initially, we aimed to map "DHB regions" to "Council regions," but this proved impractical. For social factor datasets, we filtered out those lacking DHB region information. For environment datasets, with available coordinates information, we overcame the challenge by using geographical coordinates and "R library sf" to map them to DHB geometrics.

Representing Multiple Data Points for Each Common Identifier: In the case of environment-related datasets, such as earthquakes, we simplified representation by selecting key statistical measures - maximum, average, minimum, and frequency of events - for each DHB region within a given year. The same approach was applied to datasets concerning air quality, groundwater quality, and temperature.

Creating Variables for Enhanced Representation of Social Factors: Original social factor datasets expressed counts within various categories, e.g., the number of individuals with a particular qualification level (e.g., number of people with qualification > level 3). To ensure data comparability, we converted these counts to percentages. We further analyzed the percentage of individuals with qualifications at or above specific levels, generating new variables based on cumulative sums of percentage (e.g., percentage of people with qualification > level 3). This approach was also applied to income, work hours, and birth numbers datasets.

Q5: what techniques you did use

Data wrangling: Julia is used for cancer data processing and cleaning, with packages "CSV", "DataFrames", "StringDistances", "StatsBase". R is used for risk factors data processing and cleaning, with libraries "tidyverse", "purrr", and "sf". Functions of "tidyverse" consists of *read_csv(), filter(), mutate(), select(), rename(), group_by(), pivot_wider(), pivot_longer(), summarize(), inner_join(), left_join(), lapply(), case_when(), top_n(), arrange()*. "sf" is used for mapping coordinates to DHB regions.

Visualization: R: "ggplot", "ggpubr", "ggrepel" and "ggsci" are used for generating comprehensive barplot, volcano plot, line plot and dot plot; "sf" is used for generating map plot. "ComplexHeatmap" is used for generating heatmap; "flextable" is used for generating tables. Julia: "Plots" is used for generating volcano plots.

Project management: GitHub, Git

Documentation: Rmarkdown

Q6: what you managed to achieve and what you failed to do

Achievements: We successfully collected and cleaned comprehensive cancer incidence and mortality data, as well as risk factors data. Our primary goal was to establish a connection between these diverse data sources using the "region + year" as a common identifier. This allowed us to conduct a thorough analysis of cancer incidence patterns across different regions and investigate correlations with various risk factors. The analysis revealed valuable insights and identified significant risk factors associated with the most common cancer types, including Lung, Prostate, Breast, Melanoma, and Colorectal cancers. These findings have practical implications for optimizing cancer prevention strategies.

Limitations: One notable limitation was our inability to perform subgroup analyses based on ethnics or age group, despite the potential variations in cancer incidence within different demographic groups. Additionally, our correlation analyses were conducted at the regional level, treating each region as an independent sample. Ideally, comparing risk factors among cancer and non-cancer individual could provide more accurate and robust risk factor results. Unfortunately, such individual-level data are not publicly available.