Top Mortality Causes for Republic of Korea of the years 2000, 2010, 2015, 2019*

Poisson and Negative Bionimal Modelling of Annual Death Number and Top 6 Causes

Hyuk Jang

March 16, 2024

There are various causes of deaths that are prevalent and yet we do not really care about it in our daily lives. It is crucial to note the top causes of mortality not just for governments or healthcare but also so that we citizens can work together to decrease annual number of deaths for causes that can be prevented. This paper will dive into the top mortality causes of the Republic of Korea so that we can create poisson and negative binomial models to predict which causes will have an increase or decrease in the number of annual deaths when all other conditions are kept the same.

Table of contents

*Code and data are available at: https://github.com/anggimude/Top-Mortality-Causes-of-South-Korea

1 Introduction

Death is defined as the permanent halt of all biological processes that maintain an organism. Death is inevitable to all organisms including humans, and the improvement of healthcare has allowed us to disregard many causes of deaths compared to the past. Though there still exists numerous causes to deaths in our daily lives that we don't realize that and this paper will look into the top mortality causes of the Republic of Korea. Many papers and researches tend to generalize the cause for example, instead of looking into specific cancers, they would tend to look at the annual number of deaths from all cancer types. However, we are interested in the exact causes of deaths of South Korea, not the generalized cause and its numbers.

As of 2019, the life expectancy of South Korea was 83.23 years("South Korea - Place Explorer - Data Commons" (2021)) compared to 75.91 in 2000("South Korea - Place Explorer - Data Commons" (2021)) thanks to the improvement of healthcare and technology. We see a large change in the life expectancy over the 19 years, which was an increase by 7.32 years. This brings about a question. How different would the annual number of deaths by cause change look like over the 19 year period. In 2000, Korea had a population of around 46.7 million("Population Pyramids of the World from 1950 to 2100" (2014)), and as of 2019 it was around 51.8 million("Population Pyramids of the World from 1950 to 2100" (2019)). In the 2000, most of the population was structured around the ages of $14 \sim 44$, and in 2019, it has changed to $20 \sim 64$. Overall, over the period of time we are interested in, the population structure has shifted from a young to aging society. From this information, we can expect that there would be a large increase in the number of deaths from causes that are correlated to older ages.

This paper will use poisson and negative binomial modelling to predict which causes of mortality may have an increase or decrease in annual deaths when all other conditions are unchanged. The results of the two models will be compared to check whether the results are credible to make conclusions. We want to analyze the change in demographics for all ages and gender, thus, we do not make any adjustments to the raw data obtained from WHO("Global Health Estimates: Leading Causes of Death" (2021)) in age or gender; instead we use the whole age and gender population. Initially we look into the top 10 causes of mortality for the selected years: 2000, 2010, 2015, 2019, then for further analysis and modelling we use the six common causes of mortality for all chosen years.

This paper has 4 sections in total not including the introduction. In Section 2 we look at the data that used to carry out the reports including tables and graphs of cleaned data that will be used for the models and the summary statistics. In the next section, we discuss about the models that will be used to analyze our cleaned data, how it is set up and the justifications of it. Next we display and examine the results obtained from the models including tables of the model summaries which helps us make predictions. Lastly, we make final discussions of our results and research based on each cause and dive into some weaknesses that our paper has. In addition, we explore some next steps we or anyone else interested is willing to take after reading this paper.

2 Data

2.1 Raw Data

The data used in this paper is derived from WHO("Global Health Estimates: Leading Causes of Death" (2021)) and was downloaded from the WHO Mortality Database("Global Health Estimates: Leading Causes of Death" (2021)). WHO("Global Health Estimates: Leading Causes of Death" (2021)) provides data for years country-level Global Health Estimates(GHE2019) for the years 2000-2019. Because the years that have estimates of a list provided of the cause of death categories in terms of International Classification of Diseases, Tenth Revision(ICD-10) in terms of a summary table of number of deaths by cause, age, and sex for WHO member states for the years are 2000, 2010, 2015, 2019; the data for these four years are cleaned and analyzed for this paper. The analysis of deaths by cause of the raw data is executed for the age groups from 5 to 85+. This paper looks into the data for Republic of Korea as the author is South Korean but also because WHO methods and data sources("WHO Methods and Data Sources for Country-Level Causes of Death 2000-2019" (2020)) certifies a high quality of data. The raw data includes columns such as code, cause, IS03, year, sex, age group, population, deaths, death rate per 100000 population, DALY, DALY rate per 100000 population.

The cleaning, testing, and modelling of the data for this paper was done through R (R Core Team 2023) with the aid of the following packages: tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2023), rstanarm (Goodrich et al. 2020), ggplot2 (Wickham 2016), modelsummary (Arel-Bundock 2022), bayesplot (Gabry and Mahr 2024), parameters (Lüdecke et al. 2020), broom (Robinson, Hayes, and Couch 2023), kableExtra (Zhu 2021), gt (Iannone et al. 2024), readr (Wickham, Hester, and Bryan 2024), broom.mixed (Bolker and Robinson 2022).

2.2 Cleaned Data

The data that is needed for this paper is year, cause, and number of deaths, so the raw data is cleaned to contain only the three columns we need. Because we are looking into the top 10 mortality in this paper, we rank the causes based on its number of deaths and merge it into Table 1. Now that the cleaning is done we can see the top 10 causes of deaths for the years 2000, 2010, 2015, and 2019, however, this isn't enough because to make a graph and create a poisson and negative binomial model, we must find the causes that are common in all of the years. Table 2 represents the table in which only the causes that appear among all the years descending order of ranking. Now we can recognize the six main causes of death in South Korea is stroke, Ischaemic Heart Disease, Stomach Cancer, Trachea Bronchus Lung Cancer, Liver Cancer, and Self Harm. Looking at Figure 1, we can see some interesting trends. For example, there is a plunge in the annual number of deaths from stroke over the 19 year span. On the other hand, there has been increases for the causes of Ischaemic heart disease and Trachea, Bronchus, Lung cancer. There is a slight decrease in stomach cancer while liver

cancer didn't fluctuate as much. Self harm is interesting because Korea is known to have the highest suicide rates out of all the OECD countries("Figure 2. Change in Suicide Rates, 2000 and 2011 (or Nearest Year Available)" (2013)), and we see an increase in the number deaths from self harm has increased rapidly from 2000 to 2010 and a small decrease from 2010 to 2015. It seems like the death numbers from self harm has plateaued with a very slow rate of increase. In general, the top 10 most common causes of deaths are stroke, heart disease, stomach cancer, lung cancer, road injury, liver cancer, diabetes, Cirrhosis of the liver(liver damage), and self harm.

Table 1: Top 10 Mortality Rates of South Korea

Year	Cause	Deaths	Death Rate	Ranking
2000	Stroke	44109	93	1
2000	Ischaemic heart disease	18837	39	2
2000	Stomach cancer	13205	27	3
2000	Trachea, bronchus, lung cancers	12879	27	4
2000	Road injury	12141	25	5
2000	Liver cancer	10893	22	6
2000	Diabetes mellitus	10414	21	7
2000	Cirrhosis of the liver	9968	21	8
2000	Self-harm	6860	14	9
2000	Chronic obstructive pulmonary disease	6783	14	10
2010	Stroke	31934	64	1
2010	Ischaemic heart disease	23696	47	2
2010	Trachea, bronchus, lung cancers	17121	34	3
2010	Self-harm	16852	34	4
2010	Liver cancer	12204	24	5
2010	Stomach cancer	11507	23	6
2010	Diabetes mellitus	9225	18	7
2010	Lower respiratory infections	9013	18	8
2010	Colon and rectum cancers	8886	17	9
2010	Chronic obstructive pulmonary disease	7904	15	10
2015	Stroke	28655	56	1
2015	Ischaemic heart disease	27336	53	2
2015	Trachea, bronchus, lung cancers	18806	37	3
2015	Lower respiratory infections	17164	33	4
2015	Self-harm	14255	28	5
2015	Liver cancer	12217	24	6
2015	Alzheimer disease and other dementias	11164	21	7
2015	Stomach cancer	9534	18	8
2015	Colon and rectum cancers	9405	18	9
2015	Kidney diseases	9188	18	10

Year	Cause	Deaths	Death Rate	Ranking
2019	Ischaemic heart disease	28042	54	1
2019	Lower respiratory infections	26649	52	2
2019	Stroke	25596	49	3
2019	Trachea, bronchus, lung cancers	20293	39	4
2019	Self-harm	14635	28	5
2019	Alzheimer disease and other dementias	12144	23	6
2019	Liver cancer	11589	22	7
2019	Colon and rectum cancers	10180	19	8
2019	Kidney diseases	10107	19	9
2019	Stomach cancer	8624	16	10

Table 2: Common Mortality Causes of All Four Years 2000, 2010, 2015, 2019

Year	Cause	Deaths
2000	Stroke	44109
2000	Ischaemic heart disease	18837
2000	Stomach cancer	13205
2000	Trachea, bronchus, lung cancers	12879
2000	Liver cancer	10893
2000	Self-harm	6860
2010	Stroke	31934
2010	Ischaemic heart disease	23696
2010	Trachea, bronchus, lung cancers	17121
2010	Self-harm	16852
2010	Liver cancer	12204
2010	Stomach cancer	11507
2015	Stroke	28655
2015	Ischaemic heart disease	27336
2015	Trachea, bronchus, lung cancers	18806
2015	Self-harm	14255
2015	Liver cancer	12217
2015	Stomach cancer	9534
2019	Ischaemic heart disease	28042
2019	Stroke	25596
2019	Trachea, bronchus, lung cancers	20293
2019	Self-harm	14635
2019	Liver cancer	11589
2019	Stomach cancer	8624

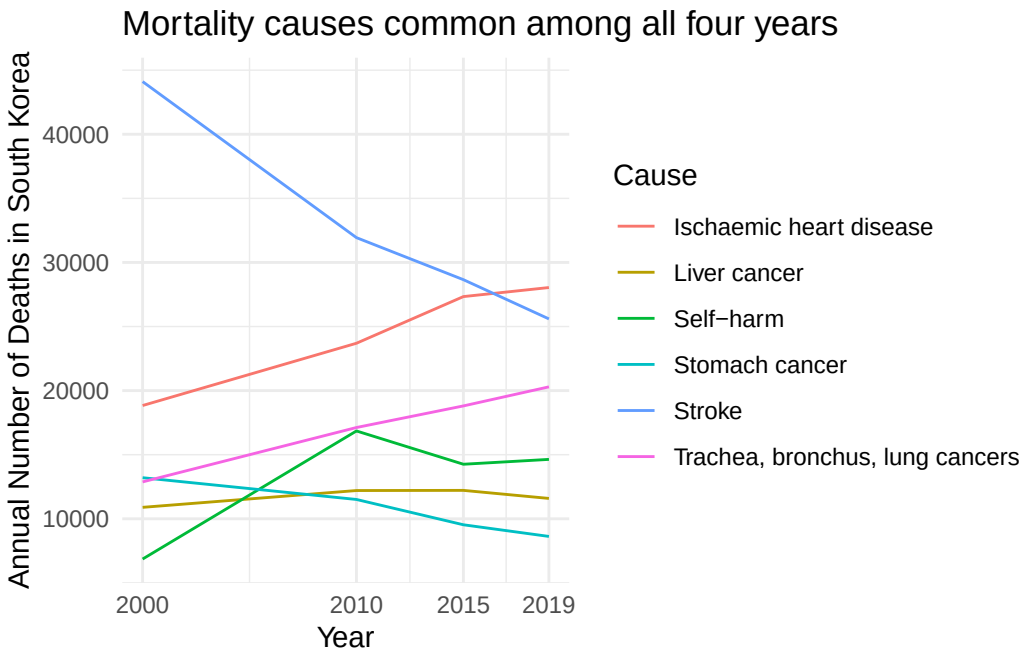


Figure 1: Mortality causes that appear in all four years

2.3 Basic Summary Statistics

Table 3 is a representation of the Table 2 showing its minimum, mean, maximum, standard deviation, variance, and sample size. The summary statistics shows that the mean of the number of deaths are 18320, with a minimum of 6860, and a maximum of 44109 for a sample size of 24 as the data selected is from Table 2. The standard deviation and variance may be abnormally high because the range of the data is large, in other words, because the mean is 18320, and the maximum is 44109, there is likely a few outliers in the data creating such a high standard deviation and variance.

Table 3: Summary statistics of the number of yearly deaths, by cause, in South Korea

	Min	Mean	Max	SD	Var	N
Deaths	6860	18 320	44 109	8927	79 687 233	24

3 Model

The goal the Bayesian model is to incorporate prior knowledge such as previous studies or analysis into the choice of model. In this paper we use poisson and negative binomial model because both of these models is often used when there occurs a certain number of events in a certain time period or intervals. The poisson distribution is efficient when used for situations where events occur independently over intervals of time. The data represents the cause and the annual number of death for the selected years, the poisson distribution is a measure that is appropriate. However, the poisson distribution is not the perfect fit because it assumes that the mean and the variance are the same whilst Table 3 shows us there is a significant difference in the mean and standard deviation. Negative binomial distribution is a perfect for situations like this. The negative binomial distribution accounts for additional variability when the data shows evidence of over dispersion. In addition, the negative binomial distribution has an extra parameter that considers the variance being larger than the mean. Thus, executing both models and comparing the results will determine the goodness of the fit.

3.1 Model set-up

The two models that are used in this paper are poisson and negative binomial which are both run in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2020). We use the default priors from `rstanarm`. Both models define y_i as the number of deaths to a specific cause. For the poisson model it is followed by the mean parameter γ_i. The mean parameter γ_i is modeled as an exponential of a linear combination of the intercept α_i, regression coefficient β_i where i is the predictor, and an extra coefficient γ_i. All three α_i, β_i, and γ_i follows a normal distribution with mean 0 and standard deviation of 2.5 as the prior. The negative Binomial model has a mean parameter μ_i and a dispersion parameter ϕ. μ_i is modeled as the exponential of a lienar combination of an intercept α_i, coefficients β_i which represents the effects of each cause on the number of deaths, and γ_i will account for any extra variability in our scenario. The intercept α_i and coefficient of cause β_i follow a normal distribution with mean 0 and a standard deviation of 2.5. The dispersion parameter ϕ follows an exponential distribution with a rate parameter of 1. In summary, both poisson and negative binomial models work towards predicting the number of deaths based on the different causes assuming each cause has a coefficient affecting the number of deaths.

$$y_i \sim \text{Poisson}(\lambda_i) \quad (1) \tag{1}$$
$$\lambda_i = \exp(\alpha + \beta_i + \gamma_i) \quad (2) \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \quad (3) \tag{3}$$
$$\beta_i \sim \text{Normal}(0, 2.5) \quad (4) \tag{4}$$
$$\gamma_i \sim \text{Normal}(0, 2.5) \quad (5) \tag{5}$$
$$\tag{6}$$

$$y_i \sim \text{NegBinomial}(\mu_i, \phi) \quad (1) \tag{7}$$
$$\mu_i = \exp(\alpha + \beta_i + \gamma_i) \quad (2) \tag{8}$$
$$\alpha \sim \text{Normal}(0, 2.5) \quad (3) \tag{9}$$
$$\beta_i \sim \text{Normal}(0, 2.5) \quad (4) \tag{10}$$
$$\gamma_i \sim \text{Normal}(0, 2.5) \quad (5) \tag{11}$$
$$\phi \sim \text{Exponential}(1) \quad (6) \tag{12}$$

3.2 Model justification

Applying the information from Table 2, it is difficult to generalize whether there would be a positive or negative correlation between the number of deaths and its cause. This is because every year there is a different number of deaths from the same cause. For example, from Figure 1, we can expect that stroke would have negative correlation coefficient because the annual number of deaths has decreased significantly over the 19 years period. On the other hand, we may predict a positive correlation coefficient for the causes like heart disease and lung cancer as they have been showing an increase trend in the number of deaths since 2000 to 2019. It may be difficult to exactly predict how self harm, liver cancer, and stomach cancer without the actual calculations being done due to the ambiguity the data shows.

4 Results

4.1 Overview of models

Our results are summarized in Table 4 and Table 5. We are primarily interested how each cause of death correlates to the annual number of deaths in South Korea. The two models provides us with th estimates for the intercept and coefficient for the causes such as liver cancer, self harm, stomach cancer, stroke, and trachea bronchus lung cancer. The intercept represents the expected log count of deaths when the cause variable is zero. However, with our data set,

the cause variable cannot be zero, which makes the interpretation of this quite unclear. The coefficients for the different causes imply the log rate ratio for each cause compared to the reference category. In other words, the coefficient shows the rate at which each cause effects the log count of deaths. Num.Obs represents the number of observations made in the model, Log.lik is the log-likelihood measuring the fitness of the model, where the lower, the better fit it is. ELPD and ELPD s.e. explains the log predictive density and its standard error. LOOIC is an acronym for leave-one-out information criterion which is a measure of model fit. Lower values are better, and the standard error shows the uncertainty of the estimate. WAIC stands for watanabe-akaike information criterion which is another measure of good fit; lower values are better fit. RMSE is the room mean squared error measuring the model's predictive performance where the lower values mean more accurate predicts.

4.2 Poisson model results

Table 4 represents the summary of the poisson model. It has an intercept of 10.105 which we can interpret as there will be 10 deaths annually from other factors. Stomach and liver cancer display a strong negative correlation coefficient each -0.826, and -0.736 respectively. Self harm also has a quite strong negative correlation coefficient -0.621. This implies that the model predicts that there will be a decrease in the log count of deaths for these causes, in short, there will be a decrease in the annual number of deaths for causes with a negative correlation coefficient when all else is held the same. Trachea, bronchus, lung(TBL) cancer has a relatively weaker negative coefficient compared to the other causes mentioned above but it is not a negligible number -0.349. Stroke is interesting because from Figure 1, we expected the coefficient to be strongly negative but it turns out to be slightly positive 0.286. For an unknown reason, the model is predicting that there will be an increase in the annual number of deaths from stroke in the near future. The poisson model has a very low log likelihood of -8157.665 implying the model may be a good fit. However, when we look at the values for LOOIC, WAIC, and RMSE, the values are extremely high each 17671.3, 22572.8, and 3830.21 which makes this model results quite unreliable.

4.3 Negative binomial model results

Table 4: Poisson model of most prevalent cause of deaths in South Korea 2000, 2010, 2015, 2019

	Poisson
Intercept	10.105
Liver Cancer	−0.736
Self Harm	−0.621
Stomach Cancer	−0.826
Stroke	0.286
TBL Cancer	−0.349
Num.Obs.	24
Log.Lik.	−8157.665
ELPD	−8835.6
ELPD s.e.	2653.7
LOOIC	17 671.3
LOOIC s.e.	5307.3
WAIC	22 572.8
RMSE	3830.21

Table 5 displays the results of the negative binomial test. Looking into the results, there is one thing different compared to what we saw from Table 4. The negative binomial model includes the standard error in parentheses which indicates uncertainty or variability and smaller standard error implies more precision. The intercept is 10.102 with a standard error of 0.190 suggesting the intercept is accurate. Liver cancer, self harm, stomach cancer, and TBL cancer is showing a negative correlation coefficient each of which is -0.726, -0.610, -0.814, and -0.314. Each has a standard error of 0.266, 0.268, 0.276, 0.272. Liver cancer, self harm, and stomach cancer exhibit the strongest negative correlation coefficient in which we can interpret as in the future, there will be a significant decrease in the number of deaths from these causes if all else is equal. TBL cancer show a relatively weaker correlation coefficient which implies that it is likely that there will be a small decrease in the number of deaths in the future. Stroke is the only cause that has a positive correlation coefficient 0.296 with a standard error of 0.274. We can say that there is a high probability that there will be a slight increase in the number of deaths caused by stroke in the future. The log likelihood is -237.44 suggesting the results of this model is accurate. The LOOIC and WAIC are relatively lower than the ones from Table 4 encourages that the negative binomial model is a better fit than the poisson model. Since the RMSE is quite similar, we can think that the negative binomial model is a better fit than the poisson model which makes the results from the negative binomial more accurate.

Table 5: Negative Binomial model of most prevalent cause of deaths in South Korea 2000, 2010, 2015, 2019

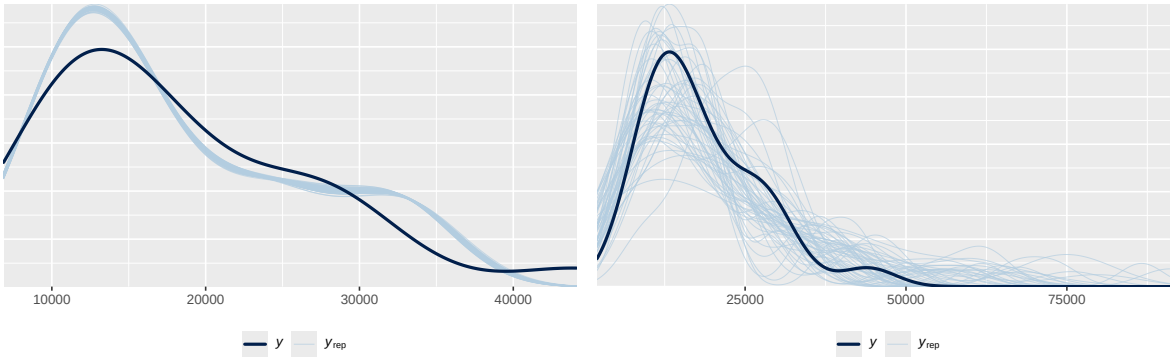
	Negative Binomial
Intercept	10.102
	(0.190)
Liver Cancer	−0.726
	(0.266)
Self Harm	−0.610
	(0.268)
Stomach Cancer	−0.814
	(0.276)
Stroke	0.296
	(0.274)
TBL Cancer	−0.341
	(0.272)
Num.Obs.	24
Log.Lik.	−237.440
ELPD	−240.6
ELPD s.e.	2.4
LOOIC	481.2
LOOIC s.e.	4.8
WAIC	481.0
RMSE	3832.16

4.4 Model results summary

Comparing the values of the two data sets, and the goodness of fit of the models, we can confirm the validity of our model results. Looking at the two values we from each model for each cause, all causes are in very close proximity. Through this we can see that the results from Table 4 and Table 5 are credible. Thus, we can confidently say that liver cancer has a negative correlation coefficient of -0.73, self harm's coefficient is around -0.61, stomach cancer is around -0.82, stroke is around 0.29, lung cancer is around -0.341 all of which have a standard error around 0.27. These results suggest that the model estimates that liver cancer, self harm, stomach cancer, and lung cancer will have a decrease in the log number of deaths while stroke will have an increase in the log number of deaths.

4.5 Other model results

Figure 2 is posterior predictive check and Figure 3 is a leave-one-out cross validation. These will be further discussed in Section A.1.



(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 2: Examining how the model fits, and is affected by, the data

```
                                    elpd_diff se_diff
cause_of_death_south_korea_neg_binomial     0.0       0.0
cause_of_death_south_korea_poisson      -8595.0    2652.1
```

Figure 3: Checking the convergence of the MCMC algorithm

5 Discussion

5.1 First discussion point

In Figure 1, we have observed that the number of deaths from stroke has decreased while lung cancer and heart disease has increased. In this section we are interested in why these trends may have occurred. Stroke is known to be a brain attack occurring when something blocks the blood supply to the brain or when a blood vessel in the brain bursts. Stroke has been known to be one of the leading causes of deaths in many countries over the past decade. This is because some can fully recover but the median survival after stoke is approximately 5 to 10 years. Over the 19 year period of our data, there has been significant improvements in healthcare, stroke rehabilitation, and advances in the acute stroke treatment globally has led to a remarkable decrease in the mortality from stroke. However, as we see in Section 4.4, the model expects an increase in the number of deaths from stroke, this could so because stroke is also caused due to stress and other unknown factors which is why despite the improvement in technology, it is still a leading factor of mortality globally.

TBL cancer is known to be the second most common cancer globally. Trachea refers to the windpipe and bronchi are the two large tubes that carry air from the trachea to each lungs. TBL cancer refers to any cancer that is caused in these areas of the human body. Smoking is known to be the most important risk factors of lung cancer and thankfully there has been a decrease in the number of smokers over time. This is maybe why the in Section 4.4, the two models expect a decrease in mortality in from TBL cancer. Then why may Figure 1 display an increasing trend. This could be due to the aging population and because of treatment delays. Lung cancer can be cured when found in its early stages, but when found in the later stages, it becomes much more difficult to cure and is likely to lead to death. Such reasons may be why the mortality from lung cancer is one of the highest even within the category of all cancer types not only any mortality causes.

5.2 Second discussion point

It is always interesting to look deep into the self harm numbers of South Korea as it is notorious around the world. One of the main reasons this trend occurs is due to the high social pressure from the competitive and success oriented society. We expect that students may have be the leading cause of this high self harm rates, but also it is because of poverty among elderly citizens. While looking at the raw data, we could observe some interesting trends between road injuries and self harm. Figure 4 represents this data and the trend. Initially in 2000, number of deaths from car injuries were twice the number of self harm, but as time passed, self harm exceeded car injuries by more than twice now. While deaths from car injuries has decreased significantly over the 19 years likely thanks to the improvement of technology, health care, and social consciousness. On the other hand, the increase in deaths from self harm is presumably from socio-economical disparities, aging population, and decreases in birth rates.

For some reason, despite the governments effort and policies, it has been possible to decrease deaths from car injuries but not from self harm. This is crucial to mention because the number of deaths from self harm is not high only among a certain gender or age group. It is not just among high school students nor elderly nor men nor female, it is the whole society that is not doing well mentally leading to such painful results. Some methods that can be implemented in decreasing the number of deaths could be by government advertisement on getting therapy. This is because in Korea, people find it difficult to talk and go to therapists and get counselling due to social pressures and negative views towards people going to therapy.

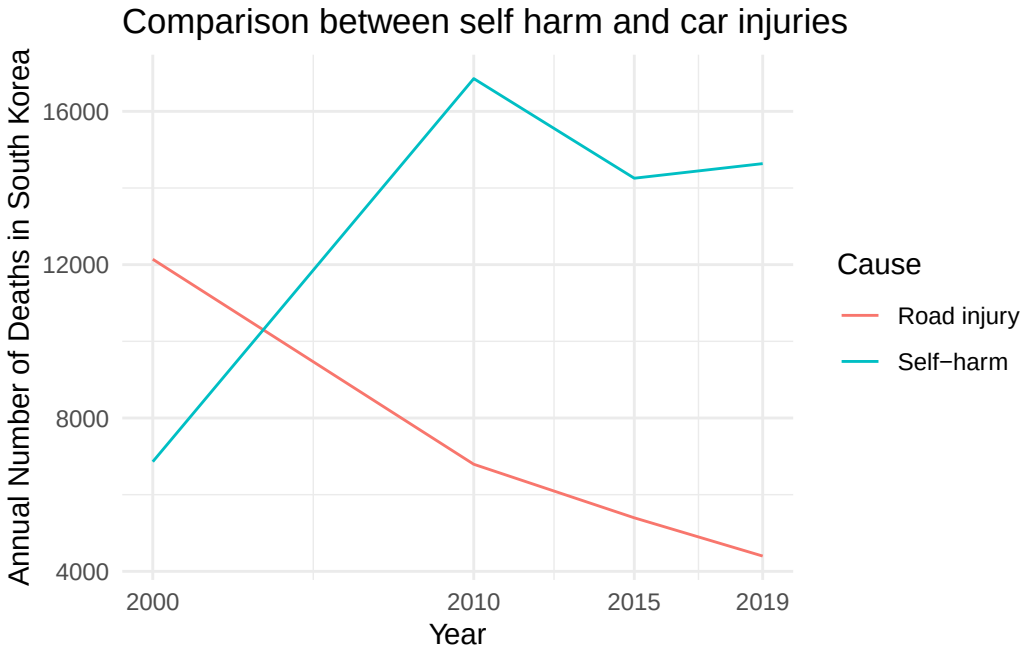


Figure 4: Comparing the sequel of car injuries and self harm numbers

5.3 Third discussion point

Going back into the initial purpose of the paper; looking into the top mortality causes of South Korea. Due to the modelling, we were only able look deep into the six common causes instead. Thus, in this section it would be more adequate to get back closer to the initial question but make it interesting as well, we look into mortality causes that appear in at least two of the four years. Figure 5 is the representation of the data that has been cleaned to analyze the new question. Now we have twelve causes to look into instead of the six we had. In addition to the ones we had there are Alzheimer disease, chronic obstructive pulmonary disease(COPD), colon and rectum cancer, diabetes, kidney disease, and lower respiratory infections. Colon and rectum cancer, diabetes, kidney disease, and Alzheimer disease do not

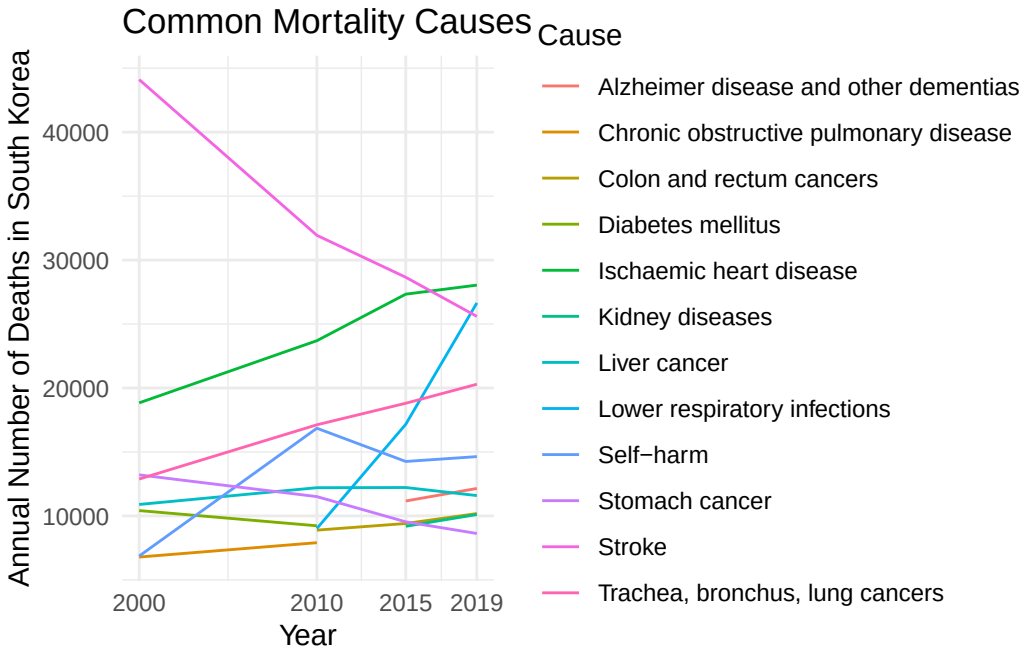


Figure 5: Mortality causes that appear in at least two years

exhibit much fluctuations. It is worth noting the steep increase in lower respiratory infections over the 9 year span. One of the main reasons of this trend is also due to aging society. Lower respiratory infection is known to have increased risk of developing when you are older. As the human body gets older, our respiratory system becomes more prone to infectious diseases. As a result, when the society is aged in general, it is likely that the number of deaths from lower respiratory infection increases as well. It is crucial to analyze and know about the top causes of mortality in countries because countries can use the data to decide where to invest in to prevent or decrease the number of people dying from certain causes so our general happiness increases from the satisfaction of life.

5.4 Weaknesses and next steps

There are some weaknesses that occur throughout the paper that may affect some of our results slightly. The number of deaths is an estimate that WHO has made based on standard categories, definitions and methods to ensure cross-national comparability. The WHO states that it is their best estimates and ensures that the quality of the data for South Korea is greatly reliable but there still may be some flaws which may have affected our analysis slightly. It may not be a significant error as the data itself is from the WHO which would ensure high quality of data however, it is worth mentioning that our raw data was based off of an estimate. Another weakness that can occur is from modelling. This is because we only looked into the

poisson and negative binomial model as we assumed these two models would be the best fit for the obtained raw and cleaned data. There may be other models such as the logistic regression or multilevel modelling which we didn't look into in this paper. Since we haven't modeled these, we are unsure as these other models could be a better fit for our data and provide us with different results. Lastly, a weakness that occurs is from modelling as well. In Figure 1 we observed six different causes of mortality in South Korea but in the Table 4, and Table 5, we only observe five of the six causes modeled. This could be due to multiple reasons such as data issue, perfect separation, model convergence, and overparametrization. Out of these four possibilities, data issue has been checked and perfect separation has been checked through box plotting the data and didn't have any issues. Thus, this weakness has likely occurred from either model convergence or overparametrization. Model convergence is wehen there is an issue in the convergence of the models, which if it is the problem, we may need to adjust the models or use other modelling techniques. Overparametrization happens when there are too many parameters to estimate the compared to the number of observations leading to a convergence issue.

For future studies, it would be suggested to use more models or more complex models to increase the accuracy of the model results and further, it would be possible to use continuous data over the most recent 10 years or all 19 years if possible. This way, we would be able to obtain much more precise data and results because this paper has used discrete data for the four years chosen; if it is continuous, we would be able to analyze trends in an explicit manner. It is crucial to do such analysis because people looking into the paper can be shocked since we normally don't see graphs like this with top causes of mortality modeled in one. Even if we do, we see it on the news for a certain cause of mortality instead all the top causes combined. This helps citizens, medical fields, and governments be more aware and decide on policies to implement to prevent or decrease the number of deaths for causes that are plausible.

Appendix

A Model details

A.1 Posterior predictive check

In Figure 2a we implement a posterior predictive check. This show the comparison between the simulated data to the actual observed data to assess whether the model is adequate. This shows the comparison for the poisson model and we can assess how well each model fits the observed data.

In Figure 2b we compare the posterior with the prior. This shows how well the negative binomial model fits the observed data.

Based on the two checks we can conclude that the negative binomial model fits the observed data better than the poisson model by visually inspecting the plots (comparing Yrep and Y).

A.2 Diagnostics

Figure 3 is a trace and Rhat plot of the negative binomial and poisson model. The trace plot shows an estimate of the model's predictive performance and helps identify any potential over fitting or under fitting issues. The Rhat plot compares the variance within individual Markov chains to the variable between the chains. A value of 1 suggests convergence, while a value greater than 1 suggests lack of convergence.

Figure 3 suggests that the negative binomial model is a much better fit for the data than poisson model we have because it has a larger ELPD value and the Rhat plot is much greater for the poisson model proposing lack of convergence.

B References

Arel-Bundock V (2022). "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software*, *103*(1), 1-23. doi:10.18637/jss.v103.i01 https://doi.org/10.18637/jss.v103.i01.

Bolker B, Robinson D (2022). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.9.4, https://cran.r-project.org/package=broom.mixed.

Bonita, R., & Beaglehole, R. (1996). The Enigma of the Decline in Stroke Deaths in the United States. Stroke, 27(3), 370–372. https://doi.org/10.1161/01.str.27.3.370

CDC. (2023, May 4). About Stroke. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/about.htm#:~:text=A%20stroke%2C%20sometimes%20called%20a,brain%20beco

Clinic, C. (2021). Bronchi: What Are They, Function, Anatomy & Conditions. Cleveland Clinic. https://my.clevelandclinic.org/health/body/21607-bronchi

Fang, M. C., Go, A. S., Chang, Y., Borowsky, L. H., Pomernacki, N. K., Udaltsova, N., & Singer, D. E. (2014). Long-term survival after ischemic stroke in patients with atrial fibrillation. Neurology, 82(12), 1033–1037. https://doi.org/10.1212/wnl.0000000000000248

Figure 2. Change in suicide rates, 2000 and 2011 (or nearest year available). (2013). https://www.oecd.org/els/health-systems/MMHC-Country-Press-Note-Korea.pdf

Gabry J, Mahr T (2024). "bayesplot: Plotting for Bayesian Models." R package version 1.11.0, https://mc-stan.org/bayesplot/.

Global health estimates: Leading causes of death. (2021). Who.int. https://www.who.int/data/gho/data/theme and-global-health-estimates/ghe-leading-causes-of-death

Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 https://mc-stan.org/rstanarm.

Ha, S. (2023, December 22). S Korea suicide: "When I found my brother's body, my heart turned to winter." Bbc.com; BBC News. https://www.bbc.com/news/world-asia-66400158

Han, K.-T., Kim, W., Song, A., Yeong Jun Ju, Choi, D.-W., & Kim, S. (2021). Is time-to-treatment associated with higher mortality in Korean elderly lung cancer patients? Health Policy, 125(8), 1047–1053. https://doi.org/10.1016/j.healthpol.2021.06.004

Hao Zhu. kableExtra: Construct Complex Table with 'kable' and Pipe Syntax, 2020. URL https://CRAN.Rproject.org/package=kableExtra. R package version 1.3.1.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J, Brevoort K (2024). gt: Easily Create Presentation-Ready Display Tables. R package version 0.10.1.9000, https://github.com/rstudio/gt, https://gt.rstudio.com.

Liu, Y., Zhang, Y., Zhao, W., Liu, X., Hu, F., & Dong, B. (2019). Pharmacotherapy of Lower Respiratory Tract Infections in Elderly—Focused on Antibiotics. Frontiers in Pharmacology, 10. https://doi.org/10.3389/fphar.2019.01237

Lüdecke D, Ben-Shachar M, Patil I, Makowski D (2020). "Extracting, Computing and Exploring the Parameters of Statistical Models using R." Journal of Open Source Software, 5(53), 2445. doi:10.21105/joss.02445.

Park, S., Choi, C.-M., Hwang, S.-S., Choi, Y.-L., Hyae Young Kim, Kim, Y.-C., Young Tae Kim, Ho Yun Lee, Si Yeol Song, & Ahn, M.-J. (2021). Lung Cancer in Korea. Journal of Thoracic Oncology, 16(12), 1988–1993. https://doi.org/10.1016/j.jtho.2021.09.007

Population Pyramids of the World from 1950 to 2100. (2014). PopulationPyramid.net. https://www.populationpyramid.net/republic-of-korea/2000/

Population Pyramids of the World from 1950 to 2100. (2019). PopulationPyramid.net. https://www.populationpyramid.net/republic-of-korea/2019/

Republic of Korea data | World Health Organization. (2020). Datadot. https://data.who.int/countries/410

Robinson D, Hayes A, Couch S (2023). broom: Convert Statistical Objects into Tidy Tibbles. https://broom.tidymodels.org/, https://github.com/tidymodels/broom.

Se Hee Jung. (2022). Stroke Rehabilitation Fact Sheet in Korea. Annals of Rehabilitation Medicine, 46(1), 1–8. https://doi.org/10.5535/arm.22001

South Korea - Place Explorer - Data Commons. (2021). Datacommons.org. https://datacommons.org/place/cou

Topic: Suicide in South Korea. (2023). Statista; Statista. https://www.statista.com/topics/8622/suicide-in-south-korea/#topicOverview

WHO. (2020). WHO methods and data sources for country-level causes of death 2000-2019. https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2019_cod_methods.pdf?sfvrsn=37bcfacc_5

Wickham H, Hester J, Bryan J (2024). readr: Read Rectangular Text Data. R package version 2.1.5, https://github.com/tidyverse/readr, https://readr.tidyverse.org.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.

Wikipedia Contributors. (2024, March 8). Suicide in South Korea. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Suicide_in_South_Korea

Wikipedia Contributors. (2024, January 15). Lower respiratory tract infection. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Lower_respiratory_tract_infection

Wikipedia Contributors. (2024, March 16). Death. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Death

Zhou, B., Zang, R., Zhang, M., Song, P., Liu, L., Bie, F., Peng, Y., Bai, G., & Gao, S. (2022). Worldwide burden and epidemiological trends of tracheal, bronchus, and lung cancer: A population-based study. EBioMedicine, 78, 103951–103951. https://doi.org/10.1016/j.ebiom.2022.103951

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models*. https://cran.r-project.org/package=broom.mixed.

"Figure 2. Change in Suicide Rates, 2000 and 2011 (or Nearest Year Available)." 2013. OECD.

Gabry, Jonah, and Tristan Mahr. 2024. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

"Global Health Estimates: Leading Causes of Death." 2021. WHO.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, JooYoung Seo, and Ken Brevoort. 2024. *Gt: Easily Create Presentation-Ready Display Tables*. https://gt.rstudio.com.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. "Extracting, Computing and Exploring the Parameters of Statistical Models Using R." *Journal of Open Source Software* 5 (53): 2445. https://doi.org/10.21105/joss.02445.

"Population Pyramids of the World from 1950 to 2100." 2014. PopulationPyramid.net.

———. 2019. PopulationPyramid.net.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*.

"South Korea - Place Explorer - Data Commons." 2021. Data Commons.

"WHO Methods and Data Sources for Country-Level Causes of Death 2000-2019." 2020. WHO.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.