Health Insurances: Gender Rating & Regional Effect on Insurance Price

Abdullah Burkan Bereketoglu
Department of Physics
Middle East Technical University
Ankara, TURKEY
burkan.bereketoglu@metu.edu.tr

Abstract

Even to this day, region and gender in many countries are believed to be one of the most important parameters to measure the pricing or cost of the health insurance that is going to apply to a person. Here in this study, the goal is to analyze causal inferences and effects of the gender and region by bayesian models built to measure the total and direct effect of gender and region. In the end, beliefs of region and gender are important parameters are discussed, and results from a conclusion are given on the case. The use of the PyMC module embedded in the Python programming language is used as the main modeling method.

Keywords: Health Insurance, Gender, Region, Causal Inference, Bayesian Statistics, Python, PyMC

Introduction

After the industrial revolution, with the increase of big cities and fast population growth, civilizations passed on a new generation of humane issues to be upheld [1]; some of these particular issues are held by private entrepreneurs with the rise of the United States. These entrepreneurs started insurance companies for certain humane matters that the government should provide by following specific government regulations and procedures that protect the consumer from the abuse of the private company. Later on, in the EU/EEA, new, more humane insurance policies and systems models were initiated[1]. Insurance, specifically health insurance, after the mid-20th century, initialized with its new model for newly emerging economic areas[1]. Countries such as Switzerland had a 100-year programmed system that continuously updated to the new age of the 21st century, and it is suggested in the Orlu et al.[1] to be adopted by Turkey due to this research to be concluded, which is gender-based insurance policies. Furthermore, another interesting pricing procedure is held in the US for people living in different states and even in different counties of a specific state, whether it is urban or rural[2]. Moreover, whether there is more than one insurance company that is giving service to the area is also effective in the current holding system in the US[2].

American Center for Disease Control and Prevention Center (CDC) suggests that particular disability and risk factors can be used like the following price policy and coverage procedure for insurances, and they can be named; Alcohol usage, Illicit Drug usage, body measurements (height, size, etc.), mandatory diet, Certain disabilities, physical or mental function issues, Exercise or physical activity capacity, Obesity/BMI, Smoking[3]. Even to date, many private and social health insurance companies still use other parameters, such as gender or age, to calculate the insurance price per year. Even today, they include different premiums for different age groups that are specially tailored to the consumer group of the specific age groups[4]. On the contrary, in recent years, gender-rating-based pricing is started to diminish by banning done by Obama legislation and European Commission, but still, many countries use both gender and age-based and other factors to calculate the insurance price[4][5]. For the US, before the ACA, which stands for the Affordable Care Act, Montana was the first to prohibit the usage of gender as an insurance price factor[5][6]. Intrinsically ACA, which was initiated and immediately put into effect in early 2010, was expected to give equal pricing and premiums for all but did not give equal pricing for all; it was missing age, income, and, most important, region[6].

Affordable Care Act started a chain-reaction in the health insurance pricing with the new regulations it came with, such that giving more potential to see the effects of the other not forbidden pricing indexes, one of which is region-based pricing[6]. Other than that, ACA did not prevent private insurance companies from switching from age to some other feature to measure their pricing for regions due to it did not forbid the usage of age[6]. With that, companies in the US started to give more importance to the region and age, even though, in recent years, there were some movements against age and income-based health insurance pricing. Furthermore, companies after the ACA even started to give numbers as ratings to the states and their counties for pricing and give different pricing for rural and urban areas in specific states according to their new policies[6][7].

In this project, the total and direct gender effect also total and direct region effect on the price will be excerpted to understand the importance and whether it is needed to be a significant factor in health insurance pricing policies. For the analysis, a data set called *Medical Cost Personal Dataset - Health Insurance Cost Prediction (Insurance) data* by Amy Aguirre is used with ~ 1300 samples[8]. Moreover, in the analysis, the dataset used, which is *Medical Cost Personal Dataset - Health Insurance Cost*

Prediction (Insurance) data, may not be a data that is taken from the United States due to its regular descriptions of the features in the metadata given by Amy Aguirre are not in English but rather Spanish, also since the Affordable Care Act, gender is no more used in the pricing or held in the data storage for the pricing reasons, so it is assumed that the data is taken from a country that has not started to use regulation such as in the US case[8].

Review

In the literature, it is seen that age and gender play a significant role in the determination of the prices of health insurance [4]. Still, many contrary regulations started to be integrated into the governmental systems against the latter, gender, playing a role in the health insurance prices and premiums. A private institution with the closing gap in premiums between the government-issued health insurance and the gender gap is now over in the EU faces a drastic decline in people who buy and continue private insurance coverages. Also, with the increasing costs, many exits the entire healthcare insurance system[14][15].

This price gap mentioned in the system is according to one of the insurance companies run in India named, Gender Rating, and still in use in India[15]. Turkey also still has this gender gap, but women after a certain age have the risk of breast cancer and other reproductive system problems[1]. By Merzel[16], it is also discussed that in low-income zones such as Central Harlem, NYC, people with low-income most of the time cannot afford private coverage, but females have the opportunity to have it covered when they are in the workforce, with that being said, females do not get affected by low-income or socioeconomic factors. On the other hand, male counterparts are heavily influenced by fewer available opportunities even in the workforce due to companies not offering them insurance making them less covered within the year 2000[16]. Literature states that certain states in the United States of America and after Obamacare now have better coverage for people from different socioeconomic backgrounds[5]. It is stated in the 2008 Los Angeles Times article that being female increases the insurance price rates drastically[17]. It is still for every one in ten women affects women between 19 to 64 age, which consists of nearly 98 million people in the United States[18]. In Taiwan, elderly or mid-aged women have different cancer insurance policies with lower claim rates for dread diseases which shows that the system is biased towards gender/age [19]. In the IFFCO-TOKIO factors article, it is stated that ten factors affect the health insurance premiums in most insurance policies. These are the recent day's age, medical history, occupation, policy duration, BMI, smoking, and location; as stated in Orlu, prices are determined by gender. Still, the premiums are not as given in the IFFCO-TOKIO article[1][20]. Lastly, Huang and Salm mentioned that the ban on gender-based pricing increased the number of people who buy health insurance, even though the new prices are slightly lower, but not drastically[21].

Another important factor to add that is not discussed in the priorly in the review is the region effect that is causing the current price disparity between whom can access how affordable health insurance and how much coverage from the premiums they can get at the end result in the United States[2]. According to Wengle[7], in the US, living in urban areas or in rural areas highly differentiates the prices and the premiums that people can have access to in the healthcare industry for insurance. For the lower competition environments, insurance companies tend to use their monopolistic power to increase the cost[7]. Hence the price is higher for the areas which have fewer health insurance companies[7]. The Affordable Care Act, or in abbreviation ACA, was initiated in the first quarter of 2010 by Barack Obama to eliminate unequal access to healthcare, eliminate the gender, smoking, and various other factors that are used in the pricing of healthcare insurance [6]. However, the region was not put in the ACA, which led companies to use the region as one of their main factors in determining new prices in different regions[2].

Methodology

In the project, it is planned to use the following approach to attack the problem of whether gender also the region individually have a tremendous effect on the prices of health insurance and also to see whether they are individually crucial indicators that should be used in the determination of price by making a directed acyclic graph of dependencies and pulling out the gender as the independent factor to affect the cost of insurance.

In the *Medical Cost Personal Dataset - Health Insurance Cost Prediction (Insurance) data*[8], there exist precisely 1338 data with a slightly imbalanced gender ratio of consumers given with a balance of 51:49 (male: female)[8]. Furthermore, according to the CIA World Factbook, the world has a 1.01:1.00 male to the female ratio, which gives the fact that the imbalance is rather insignificant when compared with the total population, so the dataset has no a bias toward gender in the light of the total population information is known[9]. The features of the data-set can be provided as; age, sex(gender), BMI, children count of the person, whether the person smokes or not, their region, and the price for the total of 6 features given[8].

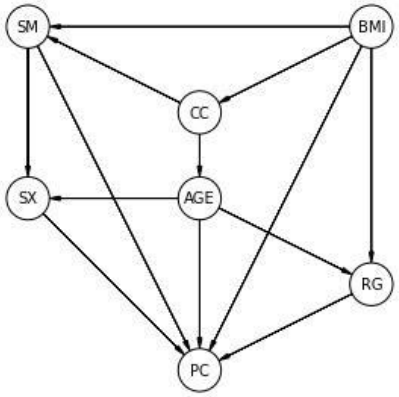

Figure 1 - Directed Acyclic Graph of Medical Cost Personal Dataset - Health Insurance Cost Prediction (Insurance) data[8] for the model to be used in the project

The dataset that is going to be used, as mentioned above, has 6 features and 1 outcome, which is the cost of the health insurance. Health insurance prices are affected by several measures, and two of them are the region and gender of the individual that is going to purchase the insurance. These two features that affect the price have an inferential domain in their system, as can be seen in *figure 1*. In some regions of a country, one can assume by their prior knowledge that in some regions there exist many elderly and in some others, many young people are living for various reasons, such as university location. Also, regions can also affect the population if small regions are picked. Technical (STEM-focused) universities and the regions near the university will have gender differences due to the lack of women in STEM fields worldwide[10]. Dataset also shows that age can also indicate gender, and it is due to the fact that natural death age imbalance of genders for people who live in equal ecosystems[11]. therefore age also has an effect on gender, and that leads to different pricing for insurance [11]. Even though the life expectancy difference is significant, it is not true to directly jump to a conclusion it is for every case for men and women living in[12]. Life expectancy measures are given for people who live similar livelihoods and end up dead at the end through a process rather than measuring for the richest women/men compared to the poorest women/men[12]. One can also infer the knowledge of whether someone is smoking or they are in older ages of their life from their children count. Furthermore, one can also infer the fact that if someone is obese, which means higher BMI, they might be quit smoking recently, or vice versa is true for the lower BMI measures. Since smoking prevalence is higher by five times in adult males, we can also indicate that sex(gender) can also be inferred from the smoking factor of the person[13]. Furthermore, in the end, *figure 1* then gives us the end-scientific model in such a way that we can find the total and direct effects of the region and gender by the abovementioned information and more.

Bayesian analysis of the data for the total and direct effects of region and sex will be conducted with the regular PyMC3 Markov Chain Monte Carlo algorithm that is used by the sampler. Furthermore, there will also be basic statistical analysis to see whether these values are correlated and give measures corresponding to the results. The PyMC3 package of Python gives us four chains of MCMC samples for our desired statistical model. The model will be created by taking the *figure 1* directed acyclic graph as the reference to build the model for direct and total effects of the desired features. For the prior statistical analysis, it is seen that before actually working on the dataset now, some correlations between various features should be stratified to find the direct effect of the gender and region. Also, find the correlation between male-female differences and smoking, smoking, and BMI. It is also will be noted that children can also be one of the factors to be taken into account in the correlation of BMI for gender regardless of male or female for the beginning; the correlation limit in this experiment will be determined by the famous economic principle called Pareto Principle which indicates that eighty-percent of the cases will be driven by twenty-percent of the indicators, so if the correlation is below twenty-percent correlation will be neglected and reported in the analysis. Later on, samples that are created by the MCMC sampling algorithm will be analyzed with diagnostic tools to measure if the sample is reliable. These measures are measures such as R-hat, trace-plots, and a number of effective samples, and various other tests will be used, such as WAIC and PSIS. In the end, we compare our four models to see which one gives better results for inferential analysis and prediction on a different basis.

In the model, the Bayesian multiple linear regression method and hierarchical analysis method will be used. The hierarchical method can be used to analyze the different regions in the system for model creation. For more intricate systems for some steps other than multiple linear regression, splines can be used after the behavior of the data in regular statistical analyses is determined. If the charge is assumed that behave logarithmically, exponentially, or in a polynomial shape, etc., a different system rather than linear regression without separating data into different sections would be more desirable. This behavior will be reported in the study if any such behavior exists in the analyses in the pre-process.

If the behavior is linear for the features for the charge, then the multiple linear regression method is to be followed. Else non-linear approaches such as splines are to be followed.

The charge predictions for gender and regions are to be visualized with summary statistics, and the highest density interval for 91 percent interval is given. 91 is picked because it is below 10 percent from tails with a slight difference. For the last step, it is to be concluded with the difference between males and females also with the regional differences to be simulated, and the differentiation statistics as the main result to be reported after specific interventions to make the gender and region independent from each other charge independently from the rest.

Results

Part A

Here stratification of the smoker parameter and age parameters in *figure 1* to see the effect of the gender parameter on the health insurance prices is done. Results show that female or male does not have a significant difference in pricing; however, prices change drastically for whether the individual is a smoker or not.

In *table 1, figure 2, figure 3, figure 4, and figure 5* one can see the result abovementioned.

Table 1: Part A ~ Gender effect on Pricing

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
a_bar	0.000	0.011	-0.020	0.019	0.000	0.000	5412.0	1476.0	1.00
alpha_A[0]	0.521	0.074	0.369	0.648	0.002	0.001	1422.0	1523.0	1.00
alpha_A[1]	0.517	0.076	0.379	0.663	0.002	0.001	1439.0	1428.0	1.00
beta_S[0]	-0.886	0.073	-1.028	-0.759	0.002	0.001	1451.0	1301.0	1.00
beta_S[1]	0.905	0.074	0.771	1.046	0.002	0.001	1280.0	1216.0	1.00
beta_A[0]	-0.008	0.010	-0.028	0.010	0.000	0.000	4244.0	1258.0	1.00
beta_A[1]	-0.007	0.009	-0.023	0.011	0.000	0.000	4853.0	1413.0	1.01
beta_A[2]	-0.004	0.010	-0.022	0.014	0.000	0.000	5603.0	1504.0	1.00

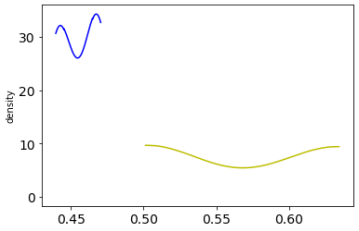

Figure 2: Density for Male~Female

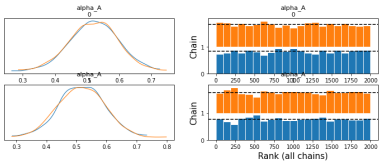

Figure 3: Rank Bars for Genders (top - female, bottom - male)

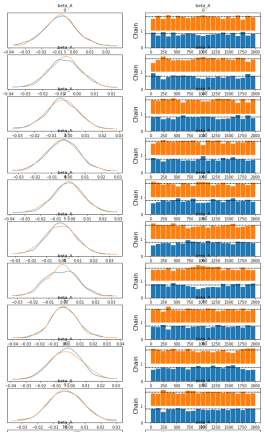

Figure 4: Rank Bars for Age on Gender model

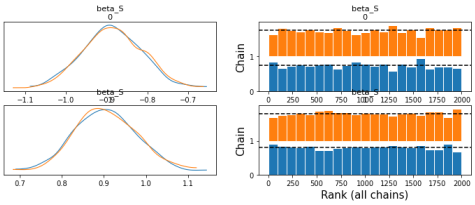

Figure 5: Rank Bars for Smokers on Gender model

Part B

Here for the region effect model, we stratified BMI and age parameters to understand the effect of the region directly and make an analysis of whether it has a significant effect on the change of health insurance prices. Results of this model show us that BMI does not have a significant effect on changing the prices; however, it adds variance to the model. Moreover, with the increasing age, it is started to be seen that prices start to have an increasing trend when compared with the younger ages, which corresponds to the first indexes of the beta_A parameter.

In *table 2, figure 6, figure 7, figure 8, and figure 9* one can see the results abovementioned in the region effect model.

Table 2: Part B ~ Region effect on Pricing

index	mean	sd
rg_bar	0.012	0.055
alpha_rg[0]	0.037	0.053
alpha_rg[1]	-0.028	0.057
alpha_rg[2]	0.093	0.056
alpha_rg[3]	-0.038	0.057
beta_bmi[0]	-0.008	0.102
beta_bmi[1]	-0.041	0.099
beta_bmi[2]	-0.015	0.101
beta_bmi[3]	-0.017	0.095
beta_bmi[4]	-0.073	0.088
beta_bmi[5]	-0.096	0.088

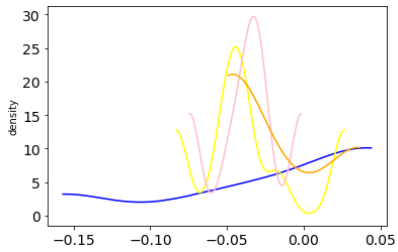

Figure 6: Density for four different regions

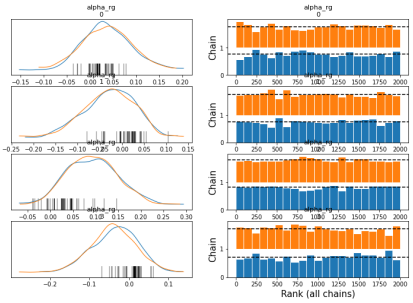

Figure 7: Rank Bars for Regions

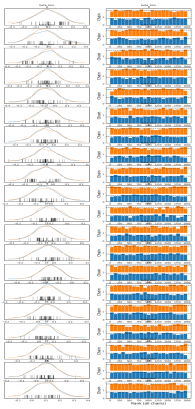

Figure 8: Rank Bars for BMI on Region model

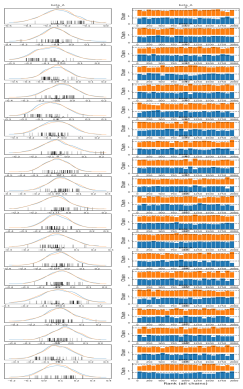

Figure 9: Rank Bars for Age on Region model

Part C

Here to see the effect of gender and region fully, by combining both of them in a multivariate Cholesky Covariance model, a different method is used.

This model suggests that region two has significantly higher health insurance prices when compared to the other three, also the first region is comparably lower than the other two in the pricing. This may be due to various reasons that will be talked about in the conclusion section. In *Table 3, figure 10, the* abovementioned results are seen.

Table 3: Part C ~ Region & Gender mixed effect on Pricing

index	mean	sd
z_A[0]	-0.711	0.687
z_A[1]	0.907	0.714
z_A[2]	-0.291	0.701
z_A[3]	-0.301	0.664
z_actor[0, 0, 0]	0.059	0.994
z_actor[0, 0, 1]	-0.016	0.998
z_actor[0, 0, 2]	0.013	1.004
z_actor[0, 0, 3]	-0.018	1.032
z_actor[0, 0, 4]	-0.018	1.016
z_actor[0, 0, 5]	-0.076	0.996
z_actor[0, 0, 6]	-0.012	0.994

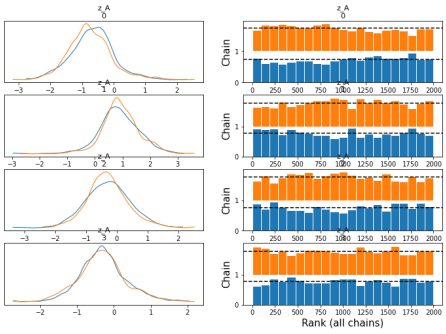

Figure 10: Rank Bars for Regions ~ Gender mixed effect

Part C - Extra

Here in this part, we continue with the model structure of the mixed effect model. Furthermore, it is seen that from the Cholesky parameter that by *table 4* and *figure 12* below, we can state that the same index Cholesky covariances are effective and gives great parametrization result.

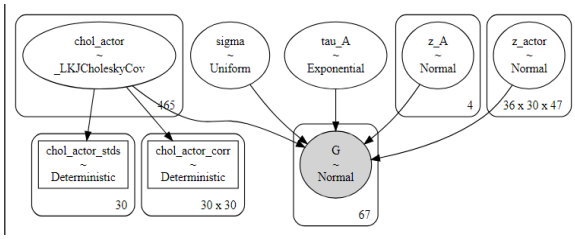

Figure 11: Mixed Effect Model

Table 4: Part C ~ Region & Gender mixed Cholesky Coefficient

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
chol_actor_corr[0, 0]	1.000	0.000	1.000	1.000	0.000	0.000	2000.0	2000.0	NaN
chol_actor_corr[0, 1]	0.002	0.167	-0.312	0.325	0.005	0.004	1147.0	1039.0	1.0
chol_actor_corr[0, 2]	0.001	0.166	-0.294	0.315	0.005	0.004	1184.0	815.0	1.0
chol_actor_corr[0, 3]	0.004	0.173	-0.330	0.316	0.007	0.005	684.0	753.0	1.0
chol_actor_corr[0, 4]	0.003	0.173	-0.308	0.325	0.005	0.004	1209.0	1206.0	1.0
...	...	...	...	...	...	...	...	...	...
chol_actor_corr[29, 25]	-0.004	0.160	-0.290	0.294	0.005	0.004	1064.0	1356.0	1.0
chol_actor_corr[29, 26]	-0.007	0.167	-0.316	0.297	0.004	0.003	1413.0	1217.0	1.0
chol_actor_corr[29, 27]	0.007	0.167	-0.305	0.323	0.005	0.003	1347.0	1186.0	1.0
chol_actor_corr[29, 28]	-0.004	0.162	-0.314	0.291	0.004	0.003	1543.0	1561.0	1.0
chol_actor_corr[29, 29]	1.000	0.000	1.000	1.000	0.000	0.000	1906.0	1940.0	1.0

900 rows × 9 columns

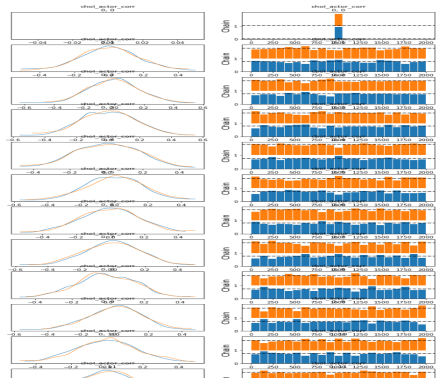

Figure 12: Cholesky Correlation

Conclusion

Here in the proposed research, to conclude the results and literature-based prior beliefs. It should be stated that gender has according to the models given above, a lower potential to have a direct effect on the pricing, positively or negatively, compared to smoking behavior, older ages, region, and comparably equal to the effect of overweight pricing based on the projections of the models built and analyzed. Even though literature suggests that gender plays a significant role, the literature, researchers never gave any background information about, whether gender, in reality, plays a significant role rather than playing a role, due to systemic bias in certain areas. It is also evident that some Western countries started to implement new regulations and insurance policies, such that disregard the effect of gender, also prohibit its usage of it. The data used may get affected by these new policies, therefore, can prove the fact that gender does not play a statistically significant role in health insurance pricing policies; however, thought as it is relevant and in the past used as a pricing index, which was as seen from the results a biased policy.

Further Work

As to study, the further studies can be concluded with higher computation power model bigger matrices with multivariate normal or more complex gaussian mixture models to analyze all the effects in a more generalized behavior, and also dive into the data used, since the US of A and some other European countries started to implement new health insurance policies that disregard gender as a pricing parameter, it can be further proved that gender is indeed not necessary to be put into consideration to give higher prices for one's self.

References

[1]Orlu, C. & Köse, A. (2018). TÜRKİYE VE İSVİÇRE GENEL SAĞLIK SİGORTASI SİSTEMLERİNİN KIYASLANMASI VE TÜRKİYE İÇİN YENİ BİR MODEL ÖNERİSİ . Finansal Araştırmalar ve Çalışmalar Dergisi , 10 (19) , 303-329 . DOI: 10.14784/marufacd.502177

[2]*Health Care Is Local: Impact of income and geography on ... - NASHP*. (2017, June). Retrieved May 21, 2022, from https://www.nashp.org/wp-content/uploads/2017/06/Health-Care-is-Local1.pdf

[3]CDC. (2016, May 12). *Disability and risk factors*. Centers for Disease Control and Prevention. Retrieved April 27, 2022, from https://www.cdc.gov/nchs/fastats/disability-and-risk-factors.htm.

[4]Yamamoto, D. H. (2013, June). *Health care costs from birth to death*. Age-Curve-Study. Retrieved April 26, 2022, from https://healthcostinstitute.org/images/pdfs/Age-Curve-Study_0.pdf

[5]Fontinelle, A. (2022, April 26). *Gender and insurance costs*. Investopedia. Retrieved April 27, 2022, from https://www.investopedia.com/gender-and-insurance-costs-5114126

[6](DCD), D. C. D. (2021, October 29). *What is the affordable care act?* HHS.gov. Retrieved May 21, 2022, from https://www.hhs.gov/answers/health-insurance-reform/what-is-the-affordable-care-act/index.html

[7]Wengle, E. (2019, September 18). *Are Marketplace premiums higher in rural than in urban areas?* RWJF. Retrieved May 21, 2022, from https://www.rwjf.org/en/library/research/2018/11/are-marketplace-premiums-higher-in-rural-than-in-urban-areas.html

[8]Aguirre, A. (2018, March 11). *Health Insurance Cost Prediction*. Kaggle. Retrieved April 27, 2022, from https://www.kaggle.com/annetxu/health-insurance-cost-prediction/

[9]Central Intelligence Agency. (n.d.). The World Factbook. Retrieved May 21, 2022, from https://www.cia.gov/the-world-factbook/countries/world/#people-and-society.

[10]*The stem gap: Women and girls in Science, Technology, Engineering, and Mathematics*. AAUW. (2022, March 3). Retrieved May 21, 2022, from.https://www.aauw.org/resources/research/the-stem-gap/.

[11]Robert H. Shmerling, M. D. (2020, June 22). *Why men often die earlier than women*. Harvard Health. Retrieved May 21, 2022, from https://www.health.harvard.edu/blog/why-men-often-die-earlier-than-women-201602199137#:~:text=I%20knew%20that%2C%20on%20average,about%207%20years%20longer%20worldwide.

[12]Crimmins, E. M., Shim, H., Zhang, Y. S., & Kim, J. K. (2019, January). *Differences between men and women in mortality and the health dimensions of the morbidity process*. Clinical chemistry. Retrieved May 21, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6345642/.

[13] World Bank Blogs. (2019, May 23). *Men smoke 5 times more than women*. World Bank Blogs. Retrieved May 21, 2022, from https://blogs.worldbank.org/opendata/men-smoke-5-times-more-women#:~:text=Smoking%20prevalence%20is%20much%20higher,daily%20or%20non%2Ddaily%20basis.

[14]Chernew, M., Cutler, D. M., & Keenan, P. S. (2005). Increasing health insurance costs and the decline in insurance coverage. *Health Services Research*, *40*(4), 1021–1039. https://doi.org/10.1111/j.1475-6773.2005.00409.x

[15]Future Generali. (2022, February 11). *How gender impacts health insurance premium*. Future Generali India Life Insurance. Retrieved April 27, 2022, from https://life.futuregenerali.in/life-insurance-made-simple/life-insurance/how-health-insurance-premium-varies-by-gender

[16]Gender differences in health care access indicators in an urban, low-income community. (2000). *American Journal of Public Health*, *90*(6), 909–916. https://doi.org/10.2105/ajph.90.6.909

[17]Lazarus, D. (2008, June 22). *Los Angeles Times: Gender can cost you in individual health*. Los Angeles Times. Retrieved April 27, 2022, from https://www.csun.edu/pubrels/clips/June08/06-23-08Y.pdf

[18]KFF. (2021, November 8). *Women's health insurance coverage*. KFF. Retrieved April 27, 2022, from https://www.kff.org/other/fact-sheet/womens-health-insurance-coverage

[19]Li, C.-S., Hung, C.-J., Peng, S.-C., & Ho, Y.-L. (2021). Impact of gender and age on claim rates of Dread Disease and cancer insurance policies in Taiwan. *International Journal of Environmental Research and Public Health*, *19*(1), 216. https://doi.org/10.3390/ijerph19010216

[20]*10 factors that affect your health insurance premium costs*. Motor, Health, Travel, and Home Insurance Online. (n.d.). Retrieved April 27, 2022, from https://www.iffcotokio.co.in/health-insurance/10-factors-that-affect-your-health-insurance-premium-costs

[21]Huang, S., & Salm, M. (2019). The effect of a ban on gender‐based pricing on risk selection in the German health insurance market. *Health Economics*, *29*(1), 3–17. https://doi.org/10.1002/hec.3958