**STA 5168:**

**Statistics in Application III**

**Yingru Liu & Ning Xue**

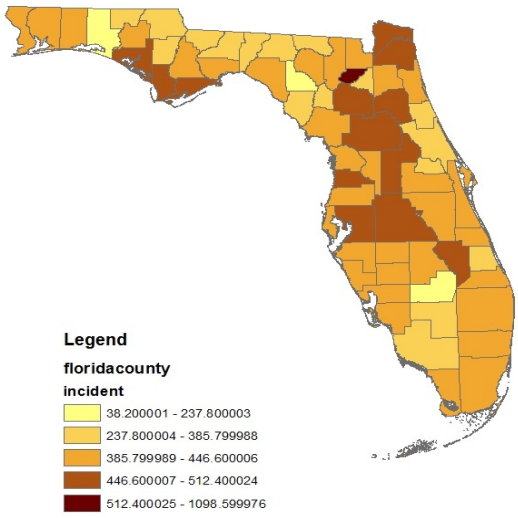
# 2010 Florida Cancer Incidence Rate and Associated Risk Factors Analysis

## BACKGROUND

### 1. Cancer spatial Disparity

Cancer is a leading cause of death worldwide. There were 7.6 million deaths (around 13% of all deaths) caused by cancer in 2008 (WHO, 2015). In America, cancer is the second most common cause of death in the US, accounting for nearly 1 of every 4 deaths (ACS, 2014). However it can be reduced and controlled by implementing evidenced-based strategies, because many cancers have a high chance of cure if detected early and treated adequately. In order to gain the achievement in cancer prevention, it is necessary to understand how cancer happens. While it is a challenge to specify one or some particular factors causing a cancer, the life style, social economic condition, environmental pollutions, genetic difference, and resistance to the mental pressure, to the environmental threats are all possible reasons accounting for cancer incidents. Cancer incidence or mortality usually appears different spatial distribution pattern although the patterns vary in different cancer types and with different environmental conditions. For example, in Hanchette's research (1992), the geographic distributions of UV radiation and prostate cancer mortality are correlated inversely ($P < 0.000$~), and prostate cancer mortality exhibits a significant north-south trend, with lower rates in the South. Many researches indicated that air pollution was highly related with lung cancer (Linder et.al., 2008; Oliveira et.al., 2013). The mortality of cancers of the lung, bladder, esophagus, stomach, large intestine, and rectum were significant associated ($p<0.002$) with 339 counties with hazard waste site (HWS) counties in the United States (Griffith, 1989). Comparing the cancer mortality spatial disparity will help to identify the major risk factors associated with cancer incidence, and further to provide preventions and surveillance strategies.

The all cancer incidence rate of 67 counties in Florida were shown on the map, seen map (a,b). From the raw data (cases per 100,000 people), 92.5% of the cancer incidence rates fell in the range of 300~500. While the highest rate, 1098.6 (Union County), was 28.8 times of the lowest rate, 38.2 (Walton County).




Map(a)

Distribution of all cancer incidence rates (per 100,000) in Florida 2010

## 2. Risk factors associated with cancer mortality

Cancers are primarily an environmental disease with 90–95% of cases attributed to environmental factors and 5–10% due to genetics(ACS, 2015). Analysis cancer spatial distribution and highly correlated factors will help identify the main influence factors. Based on the literature review, the potential risk factors could be from demographic features, social economic status, and environmental characteristics. In my research, I'm going to categorize demographic features and social status as social risk factors, and others as environmental risk factors.

### 2.1 Social Factors

Race, gender, and age are human basic biological characteristics. Cancer spatial distribution usually showed gender, racial and age disparity in most researches. While for different cancers, their spatial distribution patterns were usually different. For gastric cancer, in general, female gastric cancer incidence rates at a given age are equivalent to male rates at an age 10 years younger, also the lowest rates seen in North America and Western Europe and the highest in East Asia, South America and Eastern Europe (Forman, 2006). For bladder cancer, the rates in males were three to four times those in females (Parkin, 2008). For liver cancer of Guangxi population in China, the male-to-female mortality rate was 2.02:1; also in the same research, the total cancer death occurred mainly in the elderly population above 45 years of age, especially in people over the age of 65 (Li, 2014). For prostate cancer, the tumor stage and type varied in four population: African Americans, European Americans, Senegalese and Asian Indians (Zeigler-Johnson, 2008); Age-adjusted incidence rates in China were 2.9 per 100,000 men, while 107.8 and 185.4 per 100,000 men in white and black Americans, respectively (Klassen, 2006); in Florida, U.S.,

most counties with non-significant average annual percent change (AAPC) were located in the Florida Panhandle for white males, whereas they clustered in South-eastern Florida for black males (Goovaerts, 2011). For breast cancer, the mortality rates in black women was higher than white women, by 2005 the ratio of black to the white was 1.36 in NYC, 1.38 in the US, and 1.98 in Chicago (Whitman, 2011). For cervical cancer, in the Caucasian population, increased incidence of cervical cancer was found in a region of western coastal Finland, where frequency of two cervical cancer susceptibility genes (HLADR2 and B7) was increased, and frequency of one cervical cancer resistance gene (HLA-B15) was decreased (Castro, 2007). When focused on children and adolescents in U.S., for all cancers combined, boys had a significantly higher rate than girls, children (aged 0–14 years) had a significantly lower rate than adolescents (aged 15–19 years), and white children had the highest incidence rate among all races (Li, 2003). Overall, cancer incidence and mortality rates were higher in male and elderly group people.

Social economic status (SES) is the unique characteristics for human beings comparing with other living creatures. SES affects people in many ways. Plenty article revealed that higher cancer incidence and mortality maintained in undeveloped countries, poverty communities, and minority groups (Forman, 2006; Guay, 2014; Schootman, 2008; Hernandez, 2011; Helewa, 2013). The research of Link (1998) showed high association between use of the Pap smear and mammography; meanwhile, women with higher education and income were much likely to report being screened than those women with lower education and income. The results from Hart's research (2001) suggested that there was a significant difference between the socials classes. Lack of finance support means lack of access to medical care, to efficient medicine and medical treatments. Also the poverty limits people to obtain the higher education and get the updated information related with human health.

Occupation is another influence factor. For example, the job related with mining, chemical industries were highly related with bladder cancer (Parkin, 2008; Lopez-Abente, 2006); miners and quarrymen, farmers, fishermen, masonry and concrete workers, machine operators, nurses, food industry workers, cooks, launderers and dry cleaners might have excess risks with gastric cancer (Forman,2006). In addition to occupation, life style also plays an important role in increasing the cancer incidence. It was shown that about 30% of cancer deaths are due to the five leading behavioral and dietary risks: high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco use, and alcohol use (ACS, 2014). In the United States excess body weight is associated with the development of many types of cancer. More than half of the effect from diet is due to over nutrition rather than from eating too little healthy foods. Diets that are low in vegetables, fruits and whole grains, and high in processed or red meats are linked with a number of cancers. For example, a high-salt diet is linked to gastric cancer, and Betel nut chewing with oral cancer. Tobacco use is the most important risk factor for cancer causing 22% of global cancer deaths and 71% of global lung cancer deaths. In addition to these major factors, drinking, high mental pressure, may also relate with cancer incidence (Li, 2014).

### 2.2 Environmental factors

Industrialization and urbanization have either accurately or chronically polluted environment due to the discharge of industrial wastes in the whole world wide over the past decades. The pollutants in environment are associated with many human health problems including cancers. Air pollution is one of the major reasons that associated with lung cancer (Linder et.al., 2008;

Oliveira et.al., 2013). In the study of the cumulative cancer risk from air pollution in great Huston area, the high cancer risk areas matched most of the petrochemical complex in the Houston area very well (Linder, 2008). Oliveria et.al. (2013) studied lung cancer risk spatial characterization within dry land and dry weather events in Portugal. The results showed the dry event index had good local correlations with lung cancer. However in the further analysis, these dry land and dry weather enhanced the airborne pollution, which indicated the air pollution was highly related with lung cancer. In Mexico, The highest mortality rates due to lung cancer in both genders were observed in the north of Mexico, where located most industrialized cities (Ruız-Godoy, 2007).

Heavy metal in environment is another factor related with cancer incidence. Research has classified many heavy metals, including arsenic (As), chromium (Cr[VI]) and nickel (Ni[II]), as human carcinogens. Several researches focused on Arsenic with bladder cancer, oral cancer, and skin cancer. Arsenic pollution in mining area and farm land is very common. The high bladder mortality concentrated in mining areas in Spain was found in Lopez-Abente's research (2006), which called for attention on Arsenic. In Taiwan, central and eastern parts have very high oral cancer mortality rates, the Spatial Lag Models showed that heavy metal group, CF1 (Cr, Cu,Ni, and Zn), was most spatially related to male oral cancer mortality, which implicated that some metals in CF1 might play as promoters in OC etiology (Chiang et.al., 2010). While in Wheeler's research (2013), the association between non-melanoma cancer and environment Arsenic concentration was not obtained. It is not hard to understand the different results, since different arsenic concentration level, exposure duration and intake path varied in different researches, and these differences might associate with the different results.

The early research from Griffith (1989) might provide a clear picture how environment pollution affecting human health. The study identified 593 waste sites in 339 U.S. counties in 49 states with analytical evidence of contaminated ground drinking water providing a sole source water supply. Significant associations ($p<0.002$) between excess deaths and all HWS counties were shown for cancers of the lung, bladder, esophagus, stomach, large intestine, and rectum for white males; and for cancers of the lung, breast, bladder, stomach, large intestine, and rectum for white females when compared to all non-HWS counties. Similarly, Luo and Hendryx (2011) applied multiple linear regressions to assess the associations of carcinogenic discharges from Toxics Release Inventory (TRI) sites and lung cancer mortality rates at the county level in the United States during the years 1990 through 2007. The results revealed an excess risk of population lung cancer mortality associated with higher amounts of environmental carcinogen releases from TRI facilities in both males and females, and in both whites and African Americans. Obviously environmental pollution is highly associated with cancer mortality.

Thousands of sites in United States are contaminated due to the release of hazardous substances either from accidental spills or intentional disposal. After a potentially hazardous site is identified, the information is entered in a database, CERCLIS, and then this site is evaluated by EPA's Hazard Ranking System (HRS) and as a result by given a score. This score ranges from 0 to 100. Once the score of the site is higher than 28.5, the site is put on the National Priority List (NPL), and granted fund for cleaning up, so these sites are called superfund sites. Because most superfund sites were contaminated with carcinogens, they draw a lot attention on their health issues. The research from Budnick (1984) showed that during the 1970s, a significantly increased number of bladder cancer deaths occurred among white males in Clinton County, where had a superfund site, and a significantly increased number of other cancer deaths occurred in the general population of Clinton and three surrounding counties. In Neuberger's research (1990), it

was addressed the health problems in another superfund site, Galena, Kansas. The results showed that the environmental agents in Galena are associated with, and may have contributed to, the causation of several chronic diseases, such as chronic kidney disease, heart disease, skin cancer, and anemia in residents of this community.

### 3. Research area and goals

The research area in our project is Florida. All the data are from 2010. The goal is to explore cancer incidence rate spatial disparity in Florida and identify the associated risk factors for cancer incidence rate with statistics models. Better understanding the risk factors will help to make appropriate policies to prevent or lower the cancer incidence rates.

### 3.1 Logistic Regression

Firstly we tried Logistic Regression model, and let Y denote a binary response variable as the outcome of cancer incidence. Each observation has one of two outcomes, 1( having cancer) and 0(without cancer), which we treat as a binomial variate. Logistic regression for binary response data is one of the most important model, being commonly used in biomedical. The populations of each county will be used as the weights in our generalized linear model which is as follows.

$$\text{logit}[\pi(x)] = \quad + \quad X$$

Based on the scaled data, the logistic regression model is as follows. The odds are an exponential function of X. The odds multiply by for every 1-unit increase in X. Predictors White, Black, Bachelor, Owner-occupied Housing Rate, NonInsurance are positive associated with cancer incidence rate. We could see except predictors, Poverty and Superfund, predictors included in our model are significant. AIC score is 1331.391, and the deviation is 782.4622. When we modified the logistic regression model by deleting non-significant predictors, the model didn't become better.

| | Estimate | Std. Error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.522184 | 0.005444 | -1014.31 | < 2e-16 *** |
| White | -0.118078 | 0.019222 | -6.143 | 8.10e-10*** |
| Black | -0.090906 | 0.016980 | -5.354 | 8.62e-08 *** |
| Bachelor | -0.123271 | 0.012552 | -9.821 | < 2e-16 *** |
| Female | 0.066532 | 0.011097 | 5.995 | 2.03e-09 *** |
| older65 | 0.105231 | 0.011011 | 9.557 | < 2e-16 *** |
| Poverty | 0.007629 | 0.013913 | 0.548 | 0.583 |
| OwnHouseRate | -0.107509 | 0.010768 | -12.926 | < 2e-16 *** |

| | | | | |
|---|---|---|---|---|
| NonInsurance | -0.113394 | 0.008773 | -12.926 | < 2e-16 *** |
| MIncome | 0.085386 | 0.016097 | 5.304 | 1.13e-07 *** |
| Superfund | 0.001725 | 0.001696 | 1.017 | 0.309 |

### 3.2 Logistic GLMM

However, independent observations are assumed by GLM. But based on our data, different counties might be associated with each other from the point of geography, like race structure, education level and so on. Based on our dataset, we clustered the observations into 5 groups by clustering analysis for variables, White and Black. We extended to the generalized linear mixed model by permitting random effects as well as fixed effects. GLMM treats observations from a given group, and it has a race structure random effect for each group.

$$\text{logit}[\pi(Y_i \,|u_i\,)] = \quad + \quad X$$

From the GLMM we constructed, the AIC score and deviation both become smaller compared with logistic regression model, but the standard errors for the estimation of coefficients became relative larger than GLM.

After fitting the model, inference about fixed effects proceeds in the usual way. Asymptotic for GLMMs apply as the number of clusters increases, rather than as the numbers of observations within the clusters increase. Similarly, resampling methods such as the bootstrap using a large number of clusters should sample clusters rather than individual observations within clusters, to preserve the within-cluster dependence.

| | Estimate | Std. Error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.505234 | 0.016887 | -326.00 | < 2e-16 *** |
| Bachelor | -0.136851 | 0.012099 | -11.3 | < 2e-16 *** |
| Female | 0.066532 | 0.011097 | 5.995 | 2.03e-09 *** |
| older65 | 0.078915 | 0.009621 | 8.200 | 2.35e-16 *** |

| | | | | |
|---|---|---|---|---|
| OwnHouseRate | 0.103440 | 0.010059 | 10.300 | < 2e-16 *** |
| NonInsurance | -0.095925 | 0.007058 | -13.600 | < 2e-16 *** |
| MIncome | 0.089039 | 0.010429 | 8.500 | < 2e-16 *** |

## DATA SOURCES

In our project, the data are from three major data bases:

Cancer incidence rate (all cancer, age-adjusted) was obtained from Florida Health :

http://www.floridahealth.gov/statistics-and-data/index.html

Social economic status, demographic information and education level were obtained from USA Census Bureau, including:

Persons 65 years and over, %

Female persons, %

White alone, %

Black or African American alone, %

Asian alone, %

Hispanic or Latino,

Owner-occupied housing unit rate, %

High school graduate or higher, persons age 25 years+, %

Bachelor's degree or higher, persons age 25 years+, %

Persons  without health insurance, under age 65 years, %

Median household income

Per capita income in past 12 months

Persons in poverty, %

http://www.floridahealth.gov/statistics-and-data/index.html

Environmental pollution status ,the number of Superfund sites, was obtained from FDEP http://www.dep.state.fl.us/waste/categories/wc/pages/stat_1199.htm

R code:

```
cancer<-read.csv("~/Desktop/PROJECT/data.csv",  sep=",", header=T)

head(cancer)

cancer$InRate<-cancer$InRate/100000

cancer[,2:10]<-cancer[,2:10]/100

y <- round(cancer$InRate * cancer$Populations)

x <- scale(cancer[,2:10])

newcancer = data.frame(y, x, Superfund=cancer$Superfund,total=cancer$Populations,county=cancer$county)

head(newcancer)


#cluster analysis for education

edu <- cancer[,2:3]

dist.r <- dist(edu, method="euclidean")  #euclidean distance

heatmap(as.matrix(dist.r),labRow = F, labCol = F)

race.f <- hclust(dist.r, method="ward.D")  #Ward's method

plot(race.f) #display dendogram

groups.r <- cutree(race.f,k=3)


mds.r=cmdscale(dist.r,k=2,eig=T)

x = mds.r$points[,1]

y = mds.r$points[,2]


#install.packages("ggplot2")

library(ggplot2)

p=ggplot(data.frame(x,y),aes(x,y))

p+geom_point(size=3,alpha=0.8, aes(colour=factor(groups.r)))
```

```
p+geom_point(size=3,alpha=0.8, aes(colour=factor(groups.r)))


#glm
require(glmnet)
glm1<-glm(y/total~ x+Superfund, family=binomial, weights=total, data=newcancer)
summary(glm1)
AIC(glm1) #1331.391
glm1$dev #782.4622

glm2<- glm(y/total~ White + Black + Bachelor + Female + older65
      + OwnHouseRate + MIncome + NonInsurance, family=binomial, weights=total, data=newcancer)
summary(glm2)
AIC(glm2) #1328.709
glm2$dev #783.7807

#random effects model
library(glmmML)
fit.glmm <- glmmML(response ~ gender + z1 + z2,
            +       cluster=abortion$case, family=binomial, data=abortion,
            +       method = "ghq", n.points=70, start.sigma=9)
summary(fit.glmm)

library(lme4)
m1<-glmer(y/total~ Bachelor + Female + older65
      + OwnHouseRate + MIncome + NonInsurance +  (1 | groups.r), family=binomial, weights=total,
data=newcancer)
summary(m1)
AIC(m1) #1347.847
```

## REFERENCES CITED


1. American Cancer Society, 2014, Cancer Facts & Figures.

2. American Cancer Society, 2015, Family Cancer Syndromes.

3. Castro F.A., Haimila K., Pasanen K., et.al.. 2007. Geographic distribution of cervical cancer-associated human leucocyte antigens and cervical cancer incidence in Finland. International Journal of STD & AIDS, 18:672–679

4. Causes related with cancer, 2014, http://en.wikipedia.org/wiki/Cancer

5. Chiang C.T., Lian I.B., Su C.C., et al. 2010. Spatiotemporal Trends in Oral Cancer Mortality and Potential Risks Associated with Heavy Metal Content in Taiwan Soil. Int. J. Environ. Res. Public Health, 7:3916-3928.

6. Claeson M., Andersson E.M., Wallin M., et al.. 2012. Incidence of cutaneous melanoma in Western Sweden, 1970–2007. Melanoma Research, 22:392–398

7. Datta G.D., Glymour M.M., Kosheleva A., et al. 2012. Prostate cancer mortality and birth or adult residence in the southern United States. Cancer Causes Control, 231039-1046

8. Deng W., Long L., Li J., et al. 2014. Mortality of Major Cancers in Guangxi, China: Sex, Age and Geographical Differences from 1971 and 2005. Asian Pac J Cancer Prev,15(4):1567-1574

9. Elliott, P., & Wartenberg, D. 2004. Spatial epidemiology: current approaches and future challenges. Environmental Health Perspectives, 112(9): 998-1006

10. Forman D. 2006. Gastric cancer: global pattern of the disease and an overview of environmental risk factors. Best Practice & Research Clinical Gastroenterology, 20(4):633-649

11. Francis S.S., Selvin S., Yang W., et al. 2012. Unusual space-time patterning of the Fallon, Nevada leukemia cluster: Evidence of an infectious etiology. Chemico-Biological Interactions, 196:102–109

12. Gary G. Schwartz, Carol L. Hanchette. UV, Latitude, and Spatial Trends in Prostate Cancer Mortality: All Sunlight Is Not the Same (United States), Cancer Causes & Control, Vol. 17, No. 8, 2006, 1091-1101.

13. Gonzaga C.M.R., Freitas-Junior R., Barbaresco A.A., et al. 2013. Cervical cancer mortality trends in Brazil: 1980-2009. Cad. Saúde Pública, Rio de Janeiro, 29(3):599-608

14. Goovaerts P. and Xiao H. 2011. Geographical, temporal and racial disparities in late-stage prostate cancer incidence across Florida: A multiscale joinpoint regression analysis. International Journal of Health Geographics, 10:63

15. Griffith J., Duncan R.C., Riggan WB, et.al.. 1989. Cancer mortality in U.S. counties with hazardous waste sites and ground water pollution. Arch Environ Health, 44(2):69-74.

16. Guay B., Johnson-Obaseki S., McDonald J.T., et al. 2014. Incidence of Differentiated Thyroid Cancer by Socioeconomic Status and Urban Residence: Canada 1991–2006. Thyroid, 24(3):552-555

17. Hanchette C.L., and Schwartz G.G.. 1992. Geographic Patterns of Prostate Cancer Mortality, Evidence for a Protective Effect of Ultraviolet Radiation. Cancer, 70(12):2861-2869.

18. Hart J.. 2013. Land Elevation and Cancer Mortality in U.S. Cities Using Median Elevation Derived From Geographic Information Systems. Dose-Response: An International Journal, 11:41-48

19. Hart J.. 2011. Cancer Mortality For a Single Race in Low versus High Elevation Counties in the U.S.. Dose-Response: An International Journal, 9:348-355

20. Henry K.A., Niu X., and Boscoe F.P.. 2009. Geographic disparities in colorectal cancer survival. International Journal of Health Geographics, 8:48

21. Hernandez M.N., Chowdhury R.R., Fleming LE., et al. 2011. Colorectal cancer and socioeconomic status in Miami-Dade County: Neighborhood-level associations before and after the Welfare Reform Act. Applied Geography, 31:1019-1025

22. Klassen A.C., and Platz E.A.. 2006. What Can Geography Tell Us About Prostate Cancer? American Journal of Preventive Medicine, 30(2S):S7-S15.

23. Klassen A.C., and Platz E.A.. 2006. What Can Geography Tell Us About Prostate Cancer? American Journal of Preventive Medicine, 30(2S):s7-s15.

24. Kloog I., Haim A., Stevens R.G., et.al.. 2009. Global Co-distribution of Light at Night (LAN) and Global Cancers of Prostate, Colon, and Lung in men. Chronobiology International, 26(1):108–125

25. Li J., Thompson T.D., Miller J.W., et al.. 2008. Cancer Incidence among Children and Adolescents in the United States, 2001-2003. Pediatrics, 121(6):1470-1477

26. Linder S.H., Marko D., and Sexton K.. 2008. Cumulative Cancer Risk from Air Pollution in Houston: Disparities in Risk Burden and Social Disadvantage. Environmental Science & Technology, 242(12):4312-4322.

27. Luo J. and Michael Hendryx M.. 2011. Environmental Carcinogen Releases and Lung Cancer Mortality in Rural-Urban Areas of the United States. The Journal of Rural Health, 27:342–349

28. Michelozzi P., Barca A., Capon A., et.al.. 2002. Adult and Childhood Leukemia near a High-Power Radio Station in Rome, Italy. American Journal of Epidemiology, 155:1096–103

29. Nandakumar A., Gupta P.C., Gangadharan P., et al. 2005. Geographic pathology revisited: Development of an atlas of cancer in India. Int. J. Cancer, 116:740-754

30. Oliveira A.R., Branquinho C., Pereiral M., et.al.. 2013. Stochastic Simulation Model for the Spatial Characterization of Lung Cancer Mortality Risk and Study of Environmental Factors. Mathematical Geosciences, 45(4):437-452

31. Parkin D.M.. 2008. The global burden of urinary bladder cancer. Scandinavian Journal of Urology and Nephrology, 42(Suppl 218):12-20

32. Qureshi A.A., Laden F., Colditz G.A., et.al.. 2008. Geographic Variation and Risk of Skin Cancer in US Women, Differences between Melanoma, Squamous Cell Carcinoma, and Basal Cell Carcinoma. Arch Intern Med, 168(5):501-507

33. Ruız-Godoy L., Rios P.R., Cervantes F.S., et al. 2007. Mortality due to lung cancer in Mexico. Lung Cancer, 58:184-190

34. Schootman M., Jeffe D.B., Lian M., et al. 2008. The Role of Poverty Rate and Racial Distribution in the Geographic Clustering of Breast Cancer Survival Among Older Women: A Geographic and Multilevel Analysis. American Journal of Epidemiology, 169(5):554-561

35. Stay healthy, from ACS, http://www.cancer.org/healthy/index

36. Torres J., Correa P., Hernandez-Suarez G., et.al.. 2013. Gastric cancer incidence and mortality is associated with altitude in the mountainous regions of Pacific Latin America. Cancer Causes Control, 24:249–256

37. WheelerB.W., Kothencz g., and Pollard A.S.. 2013. Geography of non-melanoma skin cancer and ecological associations with environmental risk factors in England. British Journal of Cancer, 109: 235–241

38. Whitman S., Ansell D., Orsi J., et al. The Racial Disparity in Breast Cancer Mortality. J Community Health, 36:588–596

39. World Health Organization, 2015, Cancer.

40. Zeigler-Johnson C.M., Rennert H., Mittal R.D., et al. 2008. Evaluation of prostate cancer characteristics in four populations worldwide. Can J Urol, 15(3):4056-4064