Illinois Cancer Data Analysis

Athanasios Stamatoukos

October 6, 2018

Contents

1 Introduction

In late August 2018, the Chicago Tribune published an article regarding an EPA report which found that Sterigenics facilities in Willowbrook, IL, have been releasing ethylene oxide into the air. Ethylene oxide (EtO) is used at the Sterigenics facilities to sterilize medical equipment. It is a known carcinogen and has been linked to many different kinds of health issues. EtO is even used in creating thermobaric weapons because it is so flammable and explosive.

The two Sterigenics facilities in Willowbrook, IL, is located in a small industrial/office area which is surrounded on all sides by residences. Within 1 mile of the Sterigenics facilities, aside from the many residences, there is Gower Middle School (less than a half mile away), Hinsdale South High School (3/4 of a mile away), many parks, and popular shopping places like Target and the Willowbrook Town Center. Every day EtO is breathed in by residents, the many employees who work in the businesses immediately surrounding the facilities, the teachers and students at the nearby schools. Long term, daily exposure over the years to this harmful compound can be harmful to health.

In this analysis, I will attempt to determine whether or not the residents of the areas within a few miles of the facilities face a higher rate of cancer diagnoses than the rest of the state. It is important to note here that this analysis concerns ONLY cancer diagnoses and does not inlcude data on the many other problems linked to long term EtO exposure. The cancer data comes from the Illinois Department of Public Health and contains data of cancer diagnoses for the state from 1986-2015. For most of this analysis I use only the data from 2001-2015 because it is my understanding that the ZIP code boundaries were re-drawn in 2000 and I would not want to misrepresent my findings by using incorrect data. The ZIP codes I am using are 60527, 60439, 60561, 60521, 60558, 60514, 60559, 60525 which represent Willowbrook-Burr Ridge, Lemont-Woodridge-WilloW Springs, Darien, Hinsdale, Western Springs, Clarendon Hills, Westmont, and La Grange-Hodgkins-Countryside, respectively.

2 Analysis

I will note here that all of the code I used for this analysis is available both in the appendix of this report and in a Github repository. I performed this analysis in both R and Python. All relevant files and outputs will be in the Github repository. I want to be as transparent as possible about my methods, so for this reason I am providing all of this information. Follow this link to my repository: https://github.com/athanasios8193/illinois_cancer.

2.1 Acquiring the Data

The cancer data was taken from the Illinois Department of Public Health website (http://www.idph.state.il.us/cancer/statistics.htm). For this analysis, I downloaded the ZIP Code areas file (ZPCD8615.EXE) and the associated README (READMEv25.pdf). Each line in the file contained a single cancer diagnosis. Each diagnosis was represented as a 37 character string with each character having some kind of significance. The 'zpcd8615.dat' file has been included in my Github repository in the Data folder so you don't need to go through the process of downloading it yourself.

X1	X2
26609381 4440.7681000	-87.9886000
2361938310439.4771000	-88.3713000
22619441 4339.6049000	-87.6964000
2461944210439.6049000	-87.6964000
22629593 3337.7241000	-88.9291000
14617612 2440.5190000	-88.9819000

2.2 Cleaning and Reformating the Data

The output above shows the original format of the dataset. This is clearly not easily understood even with the instructions on how to interpret the information from the guide. My next step was to extract each portion of each line to its own column so that the data could be read more easily. I decided only to extract the values for sex, years, ZIP code, age, and type of cancer. There is also information on what stage the cancer was at during diagnosis, but I neglected to include this because I didn't deem it relevant. If anyone decides to try and replicate what I did, feel free to include that information. Whether in R or Python, the step of splitting into columns is fairly simple. The result of this process is shown in the next table.

sex	years	zip	age	type
2	6	60938	4	4
2	3	61938	4	10
2	2	61944	3	4
2	4	61944	4	10
2	2	62959	3	3
1	4	61761	4	2

This output is much nicer to look at by far, but it can still be improved upon. It is a lot of work to constantly have to refer to the guide provided by the IDPH, so I took it a step further and made reference dictionaries to replace the number codes with text to make the tables very clear. The output of this step is shown below.

sex	years	zip	age	type
female	2011-2015	60938	65+	breast invasive-female

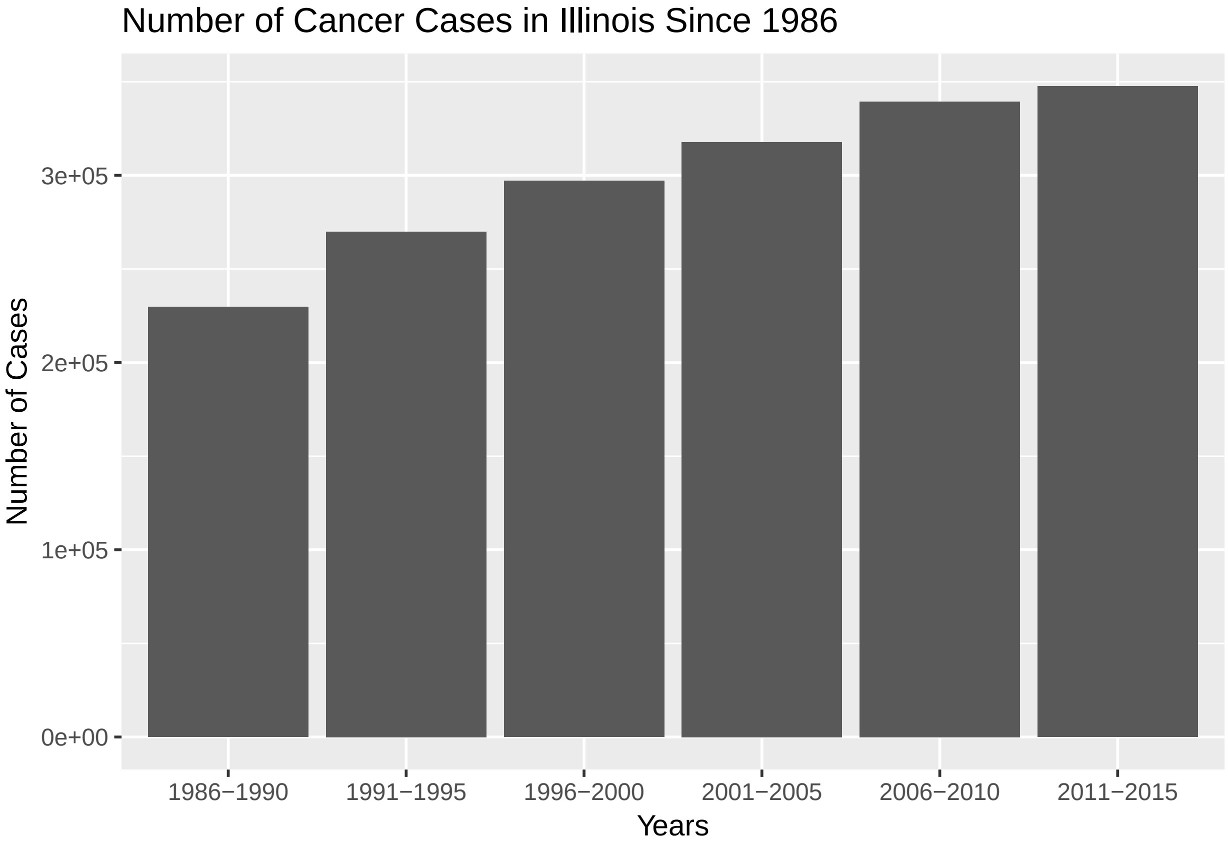
sex	years	zip	age	type
female	1996-2000	61938	65+	all other cancers
female	1991-1995	61944	45-64	breast invasive-female
female	2001-2005	61944	65+	all other cancers
female	1991-1995	62959	45-64	lung and bronchus
male	2001-2005	61761	65+	colorectal

This output is much more user-friendly than any of the others. I have included this table (18401445 rows, 5 columns) as a .csv called 'il_cancer.csv.' It is located in my Github repository in the Data folder. Feel free to explore it on your own and draw your own inferences.

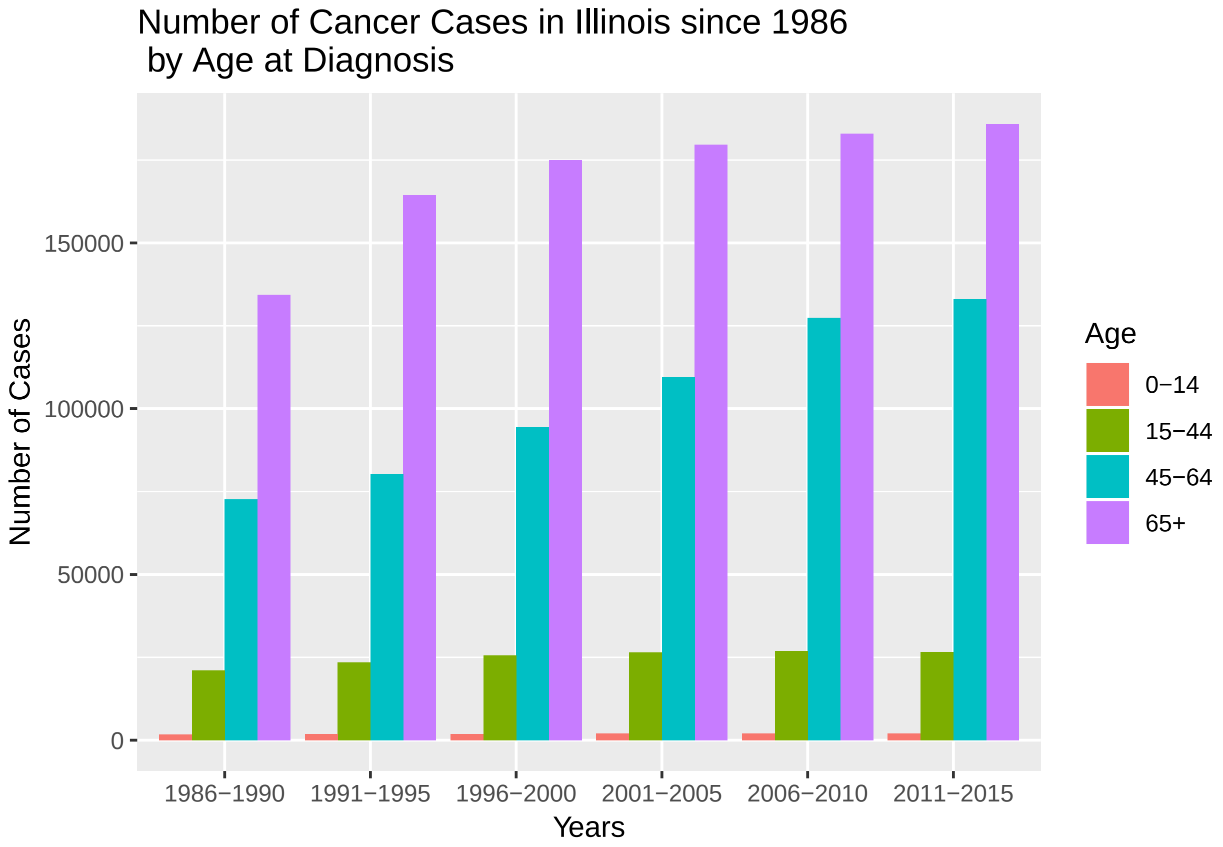
2.3 Exploring the Data

2.3.1 State of Illinois Cancer Diagnoses

Now that the data is in a clean, tidy format, it is time to explore the data. I'll start off by showing the trends for the entire state of Illinois since 1986. This first graph shows the number of cancer cases in the state as reported by the IDPH since 1986. There is a clear trend showing that for each 5-year period the number of cases goes up, but the rate of increase seems to decrease as time goes on.



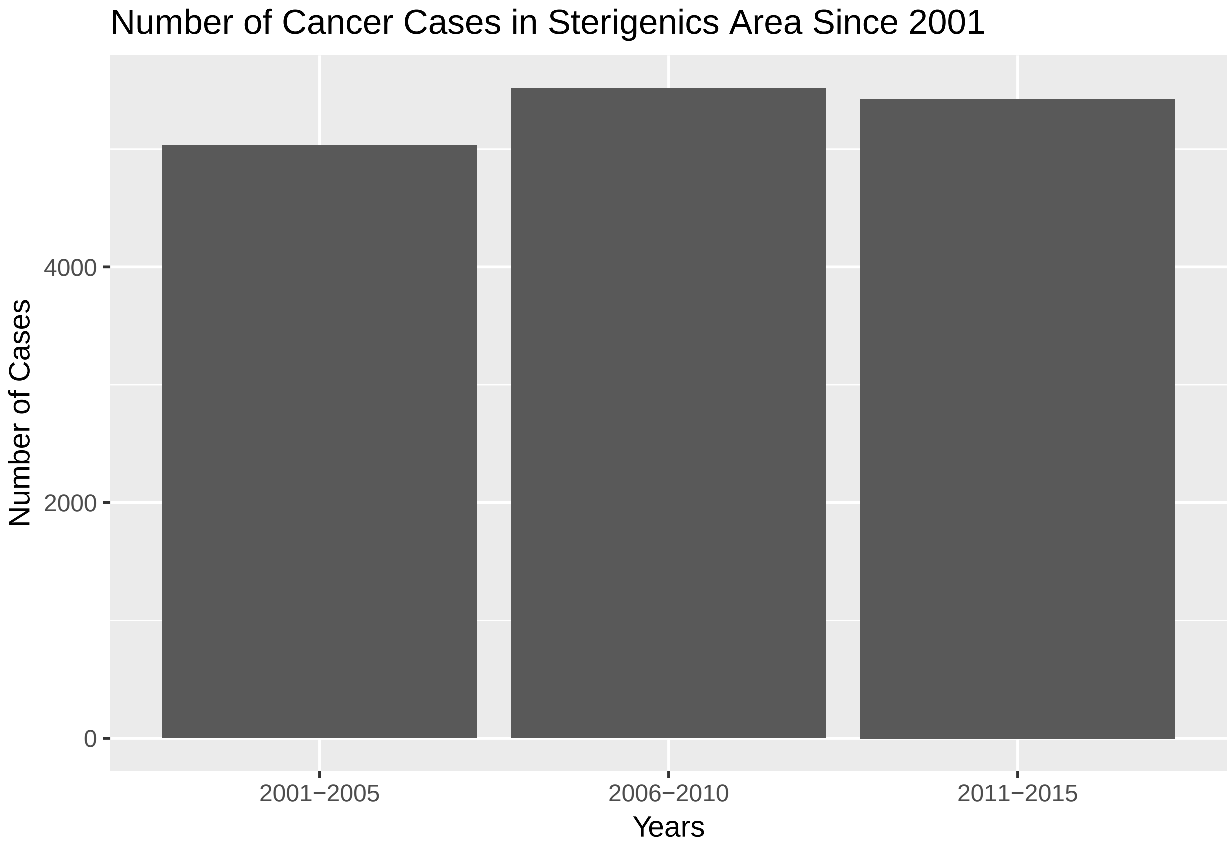
Next, I show the same graph but this time I break up each 5-year period by age. Within each age group, the number of cases goes up during each 5-year period. There is a stark difference in volume of cancer diagnoses between people above and below the age of 44. The 45-64 year age group has the highest rate of increase of cancer diagnoses over the 5-year periods. The 65 and older group is far and away the largest in terms of overall diagnoses, but the number has been fairly flat over the last 15 years in this time frame.



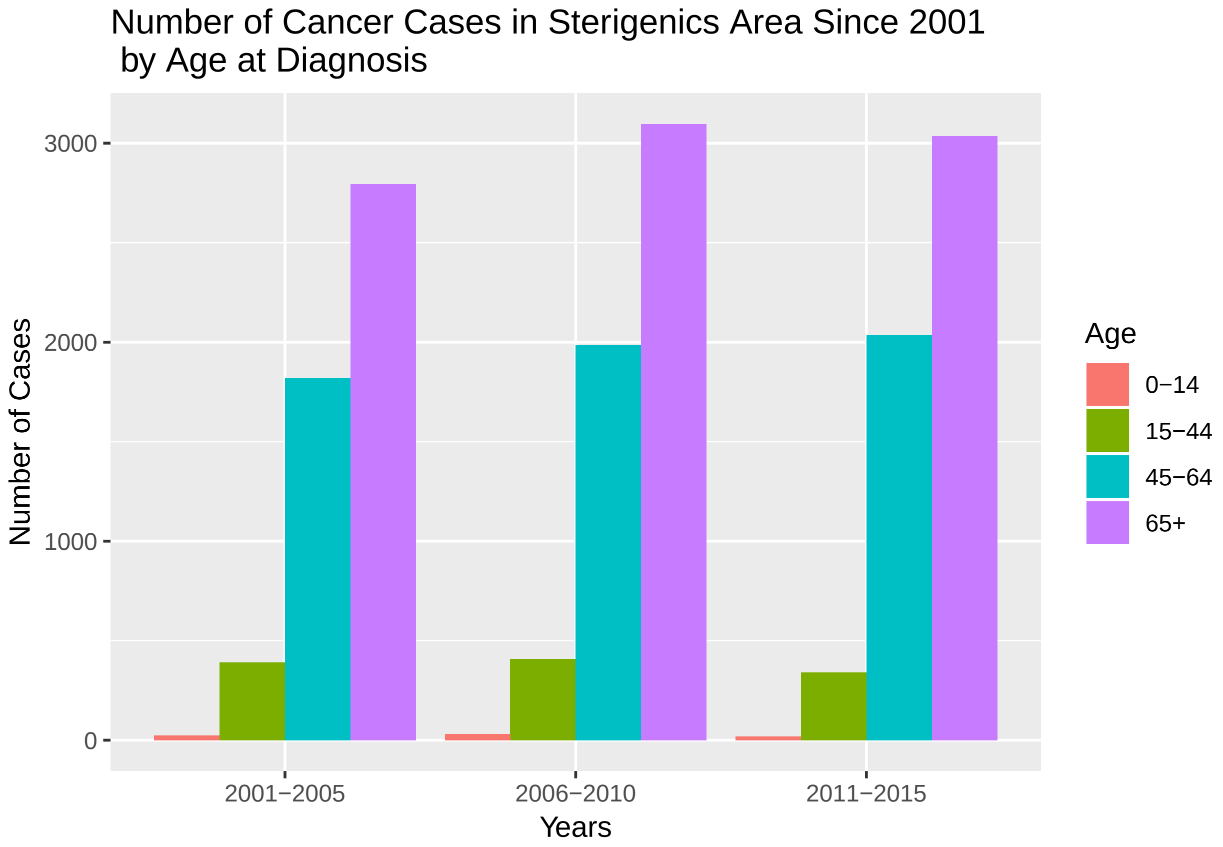
2.3.2 Cancer Diagnoses in the Areas Surrounding Sterigenics

As alluded to before, the rest of this analysis will only concern the years 2001-2015 since the ZIP code boundaries were redrawn around the beginning of that time frame. The analysis will also revolve around the areas within a few miles of the Sterigenics plants in Willowbrook to see if there is any evidence of an increased cancer rate in the area relative to the rest of the state.

From 2001-2005 to 2006-2010 the total number of cancer diagnoses went up around 500, but there was a slight decline between '06-'10 to '11-'15. Overall there were around 16,000 diagnoses in the area over that 15 year period.



When breaking up the previous graph by age, it becomes apparent that the only age group to see an increase over all three timeframes is the 45-64 group. In the other three age groups, you see an increase in total number of diagnoses form '01-'05 to '06-'10 but then a decline from '06-'10 to '11-'15.



Further exploration into the breakdown of cases between men and women or possibly to determine what kinds of cancer were most prevalent could be done, but they are beyond the scope of this high level analysis. Feel free to explore the provided data on your own.

2.3.3 Cancer Statistics Analysis

Now that overall trends have been observed, the next step is to get actual hard numbers. A good way to do this is to get normalized statistics such as '1 Cancer Diagnosis for Every X People' and 'Number of Cancer Diagnoses per 100000 People.' In order to accomplish this, I went to the U.S. Census Bureau website and got a dataset of the population of the state of Illinois per ZIP code. I used the data from the 2010 Census because it is the last year of collected data. There were estimates for more recent years as well, but I chose not to use them. Steps on downloading this dataset are in the 'cancer.R' file in the Github repository. I will also include the dataset in the Data folder of the Github repository so you don't need to repeat the steps if you want to do your own analysis.

According the the U.S. Census Bureau, in 2010 the population of the state of Illinois was 12,830,632 people.

Diagnoses_Since_2001	Cases_per_Year	One_in_Every_X_per_Year	per_100000_per_Year
1004742	66982.8	191.5511	522.0538

The results of most consequence from this table are the 'One in Every X per Year' and 'Cases per 100000 per Year,' as they are normalized numbers so comparisons can be made to other areas. These numbers say that in the state of Illinois, from 2001-2015, 1 in every 191 people was diagnosed with cancer and 522 out of every 100000 people was diagnosed with cancer. These two numbers report the exact same thing, so use whichever result makes the scale of cancer diagnoses easiest to understand.

Next, I aggregated the number of diagnoses by zip code and combined those results with the U.S. Census Data by ZIP code. I did this for every ZIP code included in the IDPH data set and calculated diagnoses per year, ‘One in Every X’ per year, and ‘X per 100000 People per Year,’ as well as the ratio of the per 100000 people per year for each ZIP code compared to the state as a whole. The first few results are shown below.

zip	population	freq	per_year	one_in_every_X_per_year	per_100000_per_year	zip_vs_state
60002	24299	1978	131.8667	184.2695	542.6835	1.0395165
60004	50582	4530	302.0000	167.4901	597.0503	1.1436567
60005	29308	2790	186.0000	157.5699	634.6390	1.2156583
60007	33820	3155	210.3333	160.7924	621.9200	1.1912948
60008	22717	1723	114.8667	197.7684	505.6419	0.9685628
60010	44095	4109	273.9333	160.9698	621.2345	1.1899817

Once I had this entire output, the results of which are saved to ‘il_cancer_statistics.csv’ (located in the Github repository’s Data folder), I extracted the ZIP codes of interest. These results are shown in the table below.

zip	population	freq	per_year	one_in_every_X_per_year	per_100000_per_year	zip_vs_state
60525	31168	3389	225.93333	137.9522	724.8888	1.3885327
60561	23115	2341	156.06667	148.1098	675.1749	1.2933052
60527	27486	2720	181.33333	151.5772	659.7298	1.2637200
60558	12960	1214	80.93333	160.1318	624.4856	1.1962093
60439	22919	2049	136.60000	167.7818	596.0120	1.1416679
60521	17597	1529	101.93333	172.6324	579.2654	1.1095895
60514	9708	817	54.46667	178.2375	561.0493	1.0746964
60559	24852	1917	127.80000	194.4601	514.2443	0.9850409

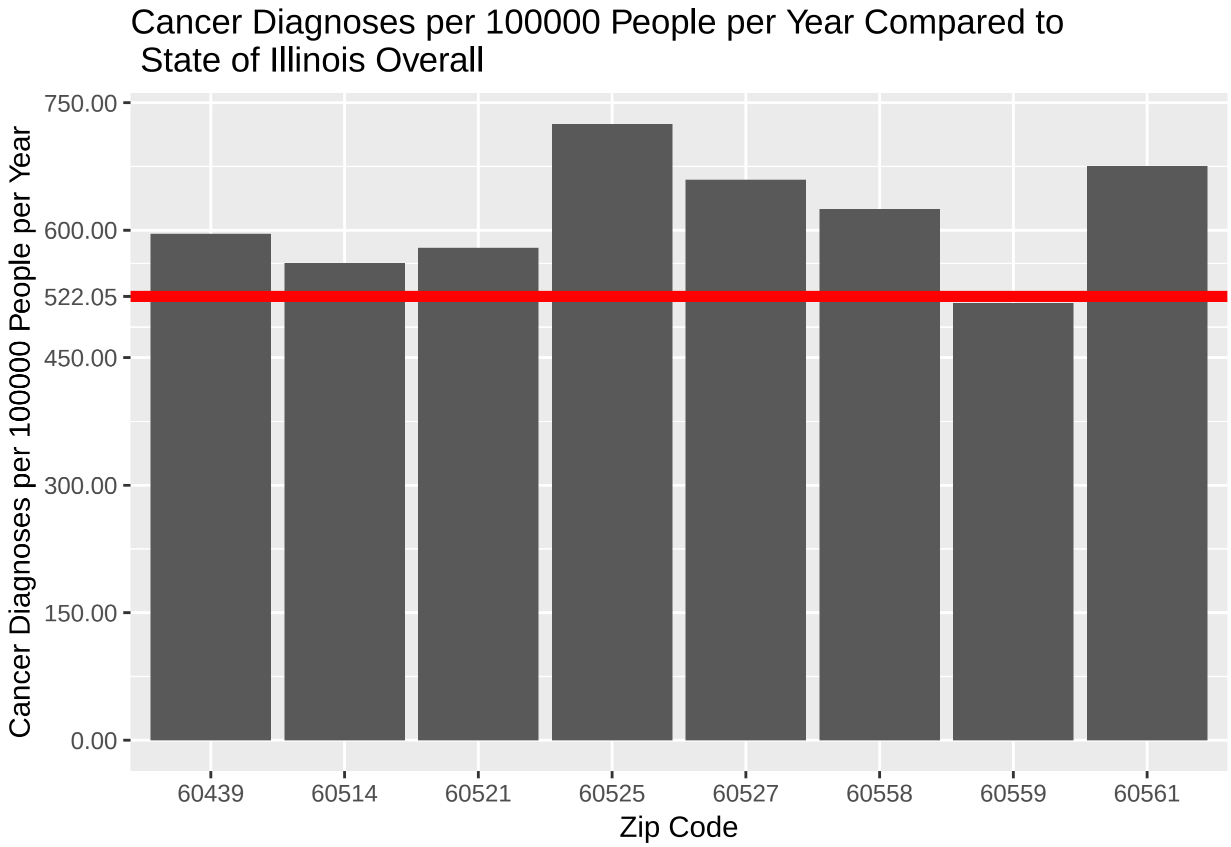
This table shows that 7/8 of the ZIP codes of interest are above the state of Illinois baseline level. La Grange, Darien, and Willowbrook are highest with 39%, 29%, and 26% more cancer diagnoses per 100000 people. Only Westmont has a lower rate than the rest of the surrounding ZIP codes.

The following table shows the proportion of ZIP codes where the Cases per 100000 People was greater than the state of Illinois overall value. I then increment the percentage higher than Illinois by 5% until I reached 20% greater. The numbers in the affected area start to get smaller due to the small sample size.

	Entire_State	Area_of_Interest
Greater_than_Illinois	0.7666185	0.875
5%_Greater_than_Illinois	0.7044798	0.875
10%_Greater_than_Illinois	0.6394509	0.750
15%_Greater_than_Illinois	0.5635838	0.500
20%_Greater_than_Illinois	0.4826590	0.375

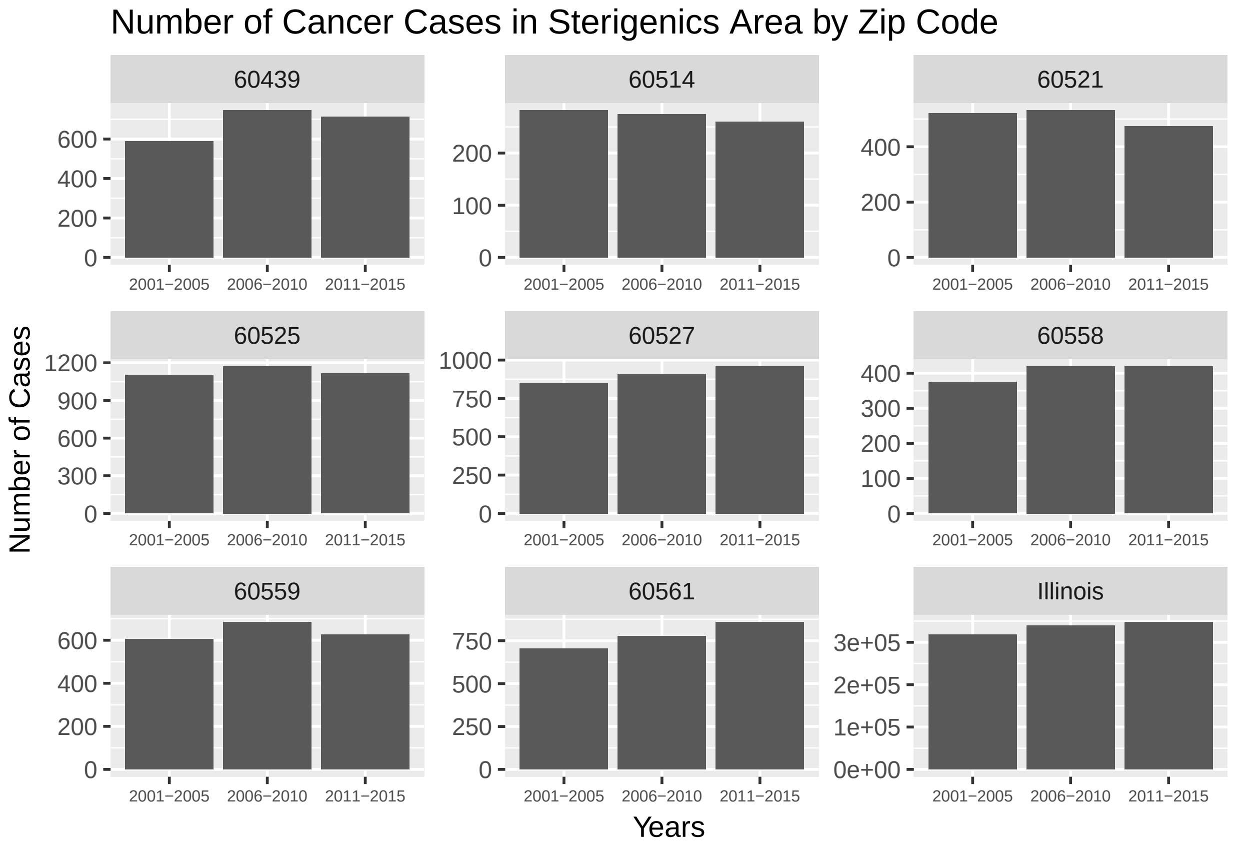
2.3.4 Graphical Comparison of Local to State

This graph is a visual representation of the per 100000 cancer rates in the Sterigenics area vs the state of Illinois as a whole (thick red line). It is clear from this that many of the local ZIP codes have a much higher rate than the state-wide baseline.



2.3.5 Local Cancer Trends Since 2001

This final graph graph shows the trend in each ZIP code in the affected area. Some ZIP codes (60527, 60558, 60561) match the state-wide trend while others defy it.



3 Conclusions

The findings above prove that the area around the Sterigenics plants in Willowbrook have a higher cancer rate than the rest of the state of Illinois. Due to the high amounts of EtO emitted from the plants, it is likely that these higher rates could be linked to the presence of Sterigenics in the community. There are many limitations to the analysis I performed. I only looked at cancer data and didn't look at other possible ailments caused by EtO. If any data exist on other diseases or illnesses caused by EtO, I could do further analysis. Also, there are many people who work in the areas surrounding Sterigenics that do not necessarily live in the area, so if they did get cancer as a result of breathing in EtO, they would be reported in their own ZIP codes and would have been excluded. I could have tried to see if the rates of breast cancer were higher in this area compared to the rest of the state as well, as there is a link between breast cancer specifically and EtO.

With all this considered, further research and testing by people more qualified than me needs to be done to determine a link between the high cancer incidence and the presence of Sterigenics.

4 Appendix

4.1 References

Illinois Cancer Data

-Illinois Department of Public Health, Illinois State Cancer Registry, public dataset, 1986-2015, data as of November 2017

Illinois Population Data

-U.S. Census Bureau; Census 2010, Summary File 1, 5-Digit ZIP Code Tabulation within Illinois; generated by Athanasios Stamatoukos; using American FactFinder; http://factfinder.census.gov; (5 October 2018)

4.2 Code

4.2.1 Loading and Cleaning Data

```r
library(dplyr)
library(ggplot2)
library(knitr)
```

```r
data <- read.delim('./Data/zpcd8615.dat', sep='\t', header=FALSE, stringsAsFactors = FALSE)
```

```r
test <- strsplit(data$V1, "   ")
mat <- t(sapply(test, '['))
data <- data.frame(mat, stringsAsFactors = FALSE)
rm(test)
rm(mat)
kable(head(data), align='c')
```

```r
data$sex <- substr(data[,1], 1,1)
data$years <- substr(data[,1], 2,2)
data$zip <- substr(data[,1], 3,7)
data$age <- substr(data[,1], 11,11)
data$type <- substr(data[,1],9,10)

data <- data[,-c(1,2)]

data$sex <- as.numeric(data$sex)
data$years <-as.numeric(data$years)
data$zip <- as.numeric(data$zip)
data$age <- as.numeric(data$age)
data$type <- as.numeric(data$type)
kable(head(data), align='c')
```

```r
sexcode <- c('male', 'female')
diagnosisyear <- c('1986-1990', '1991-1995', '1996-2000', '2001-2005', '2006-2010', '2011-2015')
agegroup <- c('0-14', '15-44', '45-64', '65+')
cancertype <- c('oral cavity and pharynx', 'colorectal', 'lung and bronchus',
                'breast invasive-female', 'cervix', 'prostate', 'urinary system',
                'central nervous system', 'lukemias and lymphomas', 'all other cancers',
                'breast in-situ-female')
```

```r
data[,1] <- sexcode[data[,1]]
data[,2] <- diagnosisyear[data[,2]]
data[,4] <- agegroup[data[,4]]
data[,5] <- cancertype[data[,5]]

rm(sexcode)
rm(diagnosisyear)
rm(agegroup)
rm(cancertype)

kable(head(data), align='c')
```

4.2.2 Exploring Data

```r
ggplot(data, aes(x=years)) + geom_bar() + xlab('Years') + ylab('Number of Cases') +
        ggtitle('Number of Cancer Cases in Illinois Since 1986')
```

```r
ggplot(data, aes(x=years, fill=age)) + geom_bar(position='dodge') +
        xlab('Years') + ylab('Number of Cases') +
        ggtitle('Number of Cancer Cases in Illinois since 1986 \n by Age at Diagnosis') +
        scale_fill_discrete(name='Age')
```

```r
since2000 <- c('2001-2005', '2006-2010', '2011-2015')
zipcodes <- c(60527, 60439, 60561, 60521, 60558, 60514, 60559, 60525)
datasub <- subset(data, years %in% since2000)
datalocal <- subset(data, zip %in% zipcodes)
datalocalsub <- subset(datasub, zip %in% zipcodes)
```

```r
ggplot(datalocalsub, aes(x=years)) + geom_bar() +
        ggtitle('Number of Cancer Cases in Sterigenics Area Since 2001') +
        xlab('Years') + ylab('Number of Cases')
```

```r
ggplot(datalocalsub, aes(x=years, fill=age)) + geom_bar(position='dodge') +
        ggtitle('Number of Cancer Cases in Sterigenics Area Since 2001 \n by Age at Diagnosis') +
        xlab('Years') + ylab('Number of Cases') +
        scale_fill_discrete(name='Age')
```

4.2.3 Comparing Cancer Rates to State of Illinois

```r
pop_illinois <- 12830632
num_cases <- nrow(datasub)
num_cases_per_year <- num_cases/15
one_in_every_il <- pop_illinois/num_cases_per_year
per_100000_il_per_year <- num_cases_per_year*100000/pop_illinois
illinois_overall <- data.frame('Diagnoses_Since_2001' = num_cases,
                               'Cases_per_Year' = num_cases_per_year,
                               'One_in_Every_X_per_Year' = one_in_every_il,
                               'per_100000_per_Year' = per_100000_il_per_year)
kable(x=illinois_overall, format='markdown', align='c')
```

```
ilpop <- read.csv('./Data/il_2010_populations.csv', header=TRUE, skip=1)
ilpop <- ilpop[,-c(1,3)]
colnames(ilpop) <- c('zip', 'population')
```

```
ilcancer <- datasub %>% count(zip)
ilcancer <- as.data.frame(ilcancer)
colnames(ilcancer) <- c('zip', 'freq')
```

```
cancer <- left_join(ilpop, ilcancer)
cancer$per_year <- cancer$freq/15
cancer$one_in_every_X_per_year <- cancer$population/cancer$per_year
cancer$per_100000_per_year <- cancer$per_year*100000/cancer$population
cancer$zip_vs_state <- cancer$per_100000_per_year/per_100000_il_per_year
kable(head(cancer), align='c')
```

```
cancer_local <- subset(cancer, zip %in% zipcodes)
kable(arrange(cancer_local, desc(zip_vs_state)), align='c')
```

```
cancer_greater <- c(nrow(subset(cancer, zip_vs_state > 1))/nrow(cancer),
                    nrow(subset(cancer, zip_vs_state > 1.05))/nrow(cancer),
                    nrow(subset(cancer, zip_vs_state > 1.10))/nrow(cancer),
                    nrow(subset(cancer, zip_vs_state > 1.15))/nrow(cancer),
                    nrow(subset(cancer, zip_vs_state > 1.20))/nrow(cancer))
cancer_local_greater <- c(nrow(subset(cancer_local, zip_vs_state >1))/nrow(cancer_local),
                          nrow(subset(cancer_local, zip_vs_state >1.05))/nrow(cancer_local),
                          nrow(subset(cancer_local, zip_vs_state >1.10))/nrow(cancer_local),
                          nrow(subset(cancer_local, zip_vs_state >1.15))/nrow(cancer_local),
                          nrow(subset(cancer_local, zip_vs_state >1.20))/nrow(cancer_local))
cancer_greater_both <- data.frame('Entire_State'=cancer_greater,
                                  'Area_of_Interest'=cancer_local_greater,
                                  row.names = c('Greater_than_Illinois',
                                                '5%_Greater_than_Illinois',
                                                '10%_Greater_than_Illinois',
                                                '15%_Greater_than_Illinois',
                                                '20%_Greater_than_Illinois'))
kable(cancer_greater_both, align='c')
```

```
ggplot(cancer_local, aes(x=as.factor(zip), y=per_100000_per_year)) + geom_bar(stat='identity') +
        geom_hline(yintercept = per_100000_il_per_year, lwd=2, color='red') +
        xlab('Zip Code') + ylab('Cancer Diagnoses per 100000 People per Year') +
        scale_y_continuous(breaks = sort(c(seq(0, 750, length.out=6),
                                           round(per_100000_il_per_year,2)))) +
        ggtitle('Cancer Diagnoses per 100000 People per Year Compared to \n State of Illinois Overall')
```

```
local_counts <- datalocalsub %>% group_by(years) %>% count(zip)
local_counts$zip <- as.character(local_counts$zip)

illinois_counts <- datasub %>% group_by(years) %>% count()
illinois_counts$zip <- c('Illinois', 'Illinois', 'Illinois')
illinois_counts <- illinois_counts[,c(1,3,2)]
```

```r
counts <- rbind(local_counts, illinois_counts)

ggplot(counts, aes(x=years, y=n)) + geom_bar(stat='identity') +
        facet_wrap(.~zip, scales='free') + theme(axis.text.x=element_text(size=6)) +
        ggtitle('Number of Cancer Cases in Sterigenics Area by Zip Code') +
        xlab('Years') + ylab('Number of Cases')
```