The Use of Geographic Information Systems in Analyzing the Spatial Distribution of People at Risk for Thyroid Cancer

Kelly M. Fox
Department of Resource Analysis, Saint Mary's University of Minnesota, Winona, MN 55987

Keywords: Thyroid Cancer, GIS, Geographic Information Systems, Incidence, Risk Factors, Spatial Analyst

Abstract

An increase in thyroid cancer incidence rates in the past decade has recently brought this disease to public attention. Unfortunately, much about the nature of this disease is unknown. This project used thyroid cancer incidence data from the National Cancer Institute and compared it with a risk factor analysis, completed using the Spatial Analyst extension in ESRI's ArcMap software. In addition, this risk factor analysis shows how Geographic Information Systems (GIS) can be an important tool in the analysis of this disease. The risk factor analysis used in this comparison identified at-risk populations based on the commonly recognized risk factors of radiation, gender, age and race. A statistical analysis of these two datasets found that there was no significant linear correlation between a risk factor analysis and incidence rate. However, it was able to provide some important information that was useful in future analyses. When the incidence rates and risk factor analysis data were spatially compared, the West and Midwest were found to have the largest difference. These results suggest that future analysis should be focused in these areas to find which risk factors play a smaller or larger role in incidence rates. Eventually, this information could help researchers identify factors that seem to have the largest affect on thyroid cancer to help people most at-risk for getting this disease by allowing them to obtain the information, treatment, and hopefully the proactive prevention methods they need.

Introduction

A study conducted by the National Cancer Institute using the Surveillance Epidemiology and End Results program found a 6.7% increase in thyroid cancer (with a 95 % confidence interval between 4.0 and 9.6) between the years 2000 and 2004 (National Cancer Institute, 2007a). This has led to much debate as to why this increase is occurring.

An article in the Journal of the American Medical Association (JAMA) suggests two possible sources for this increase 1) an actual increase in cancer incidence due to some unknown factor or 2) an increase in detection of small cancers due to more accurate detection methods (Davies and Welch, 2006).

It was beyond the scope of this project to determine the cause of this rise in incidence rates. Instead, this project used Geographic Information Systems (GIS) to identify the location of at-risk populations so that information can be given to the people that need it. Moreover, it presented a method of

Fox, Kelly. 2008. The Use of Geographic Information Systems in Analyzing the Spatial Distribution of People at Risk for Thyroid Cancer. Volume 10, Papers in Resource Analysis. 10pp. Saint Mary's University of Minnesota Central Services Press. Winona, MN. Retrieved (date) from http://www.gis.smumn.edu

studying risk factors, or potential risk factors, spatially. With further research, this information could be used to more thoroughly understand how risk factors affect the disease. Eventually, this may lead to a better explanation for the rise in incidence rates.

Some of the data used for this analysis, while publicly available, is sensitive and no attempt should be made to use this data or the analysis provided in this paper to determine the identity of any person or establishment involved in this study.

Background

There are five main types of thyroid cancer: papillary cancer, follicular cancer, medullary cancer, anaplastic cancer, and thyroid lymphoma (Mayo Clinic Staff, 2007). This study focused on the two most common types of thyroid cancers: papillary cancer (papillary carcinoma, papillary adenocarcinoma) and follicular cancer (follicular carcinoma, follicular adenocarcinoma) accounting for 90% of thyroid cancers (Mayo Clinic Staff, 2007).

A major component of creating a model that accurately represented a population's risk for these thyroid cancers was identifying common risk factors. These risk factors were characteristics that independently increased a person's risk of getting the disease. It is important to note, however, that having one or more of these factors does not necessarily mean an individual will get this disease. Conversely, lacking any, or all, of these factors does not mean that one is entirely safe (American Cancer Society, 2005).

A review of literature on thyroid cancers identified 6 major risk factors: radiation, gender, low iodine diet, age, race and family history. This analysis focused on only 4 of these factors: radiation, gender, age and race. A low iodine diet was ignored because it is not usually a problem in the United States where iodized salt has been added to many of the foods served here (American Cancer Society, 2005).

Family history was ignored for two reasons. First, family history is often not spatially distributed. Second, it would be difficult, because of privacy issues, to obtain enough data for the scale of this project.

Radiation

Exposure to high levels of radiation, specifically Iodine-131 (I-131), has been found to greatly increase a person's chances of getting papillary or follicular thyroid cancer (National Institute, 2002).

In the United States, one of the most likely sources of this radiation is from nuclear testing in the late 1950's in Nevada (National Cancer Institute, 2007b). Children under the age of 15 at the time of these tests who drank a significant amount of milk were particularly susceptible. While much of the radiation in the air was dispersed, the radioactive iodine in grass built up in cow's milk and when consumed by humans was then absorbed by thyroid glands, potentially leading to thyroid cancer (National Cancer Institute, 2007b).

In terms of at-risk populations, this study identified areas with high levels of I-131 radiation from these nuclear tests with a large percentage of the population born before 1971. Although there were other forms of radiation exposure that may have affected the population, they either could

not be easily mapped (people who used x-rays) or the risk in the United State was deemed too small to warrant inclusion (fallout from the disaster at Chernobyl) (National Cancer Institute, 2007b).

Gender and Pregnancy

Studies have shown that women are 3 times more likely than men to get thyroid cancer (Mayo Clinic Staff, 2007). There also has been evidence that pregnancy, especially later in life, could affect one's risk of getting the disease (Mayo Clinic Staff, 2007). For this analysis, women who had a baby within the last 12 months were considered at higher risk than men or women who were not pregnant within this time period.

Age

Although thyroid cancer has been found in people of all ages, the two types of cancer that are the focus of this analysis occur mainly in people from ages 20 to 60 (American Cancer Society, 2005).

Race

There is also evidence that white Americans are more susceptible to this form of cancer than black Americans (Mayo Clinic Staff, 2007). This means areas with a lower relative black population were considered at a higher risk.

Methods

Software Requirements

The GIS analysis of this project involved using ArcGIS 9.2 ArcMap and ArcCatalog with an ArcInfo license and the Spatial Analyst extension. In addition, Microsoft Excel 2007 was used with the data obtained from the American FactFinder website to convert these tables into comma delimited files.

Data Collection and Manipulation

This analysis used data from a variety of sources. Base layer state and county datasets were obtained from the datasets available with the ArcGIS software.

The state dataset was then joined with 2004 Community Survey tables on gender, pregnancy, age and race obtained from the US Census American FactFinder website and converted from pipe delimited files to comma-delimited files using Microsoft Excel.

For gender and age, the 2006 'Age by Sex' table was used with the field calculator to calculate percents of the population of a particular type. For gender, the percent female was found by dividing the field 'Estimated Total Population Female' by the 'Estimated Total Population' field.

Meanwhile, the percent of the total population between the ages 20 to 59 was found by adding together the 'Ages 20 to 59 Male' and the '20 to 59 Female' fields and the dividing this sum by the 'Estimated Total Population' field.

The pregnancy values, from the table 'Martial Status by Age' in the American Factfinder, were found by using the field calculator to divide the 'Women 15 to 50 years: Number of women who had a birth in the past 12 months (Estimate)' field by the 'Total Population of women 15 to 50 years old'.

Race data was found using the 'Race' table from the American Factfinder 2004 Community Survey. To

get usable values the 'Black' field was divided by the 'Total Population' field in field calculator on a new field. Unlike the preceding tables, the higher values were considered a lower risk as black Americans are found to be at a lesser risk than other races.

Radiation data was obtained from the National Cancer Institute's State and County Exposure Levels to I-131 and was manually input into the county dataset (National Cancer Institute, 2007a).

Finally, cancer incidence data was exported into a comma delimited file from the 'Latest incidence rates for the United States Thyroid, All Races' 2004 table from the National Cancer Profiles on the National Cancer Institute website (National Cancer Institute, 2007a).

Rasterization and Reclassification

A major part of the preparation for analysis was getting the risk factor datasets in a format that allowed them to be combined, and eventually compared, with the cancer incidence dataset. To accomplish this task, the 5 risk factor datasets (gender, age, race, pregnancy, and radiation) had to be converted into a raster format. This was completed by using the Convert > Features to Rasters tool in the Spatial Analyst extension of ArcMap. The large scale and nature of this project rendered the output cell size of these rasters unimportant; however, to stay consistent, a cell size of 0.10 decimal degrees, or about 8.4 miles, was chosen for all 5 of the newly created rasters.

Since the radiation data was obtained on a county level scale rather than a state level scale, it had to be converted to the state scale to make it consistent with the other risk factors. To accomplish this, the Zonal Statistics tool in ArcMap was used with the newly created raster data to find the mean radiation level for each state. This tool has the ability to calculate statistics (in this case the mean values) of the raster layer within a particular zone (in this case zones were defined by state).

Once the mean radiation level was determined, it was added to the states table, which was then used to create another raster based on this field.

After the rasters were created, each of them had to be reclassified so that they could be more easily compared. This was done by classifying the existing data into 10 classes. For age, race, pregnancy and gender this classification was completed by using Jenk's natural breaks as used also by Slocum et al. (2004). Since these data were all ratios of the total population rather than specific values of the risk factor, this method seemed to best illustrate the differences between the states.

Radiation, unlike the other risk factors, was divided by specific values of rads. This meant that when classifying the values it made more sense to input values of significance. For this analysis these values were reclassified into the following classifications:

< 0.519670	1
0.519670 - 0.750000	2
0.7500001 - 1.000000	3
1.000001 - 2.000000	4
2.000001 - 3.000000	5
3.000001 - 4.000000	6
4.000001 - 5.000000	7
5.000001 - 6.000000	8
6.000001 - 7.000000	9
7.000001 - 8.000000	10

Once all the data were classified,

they were then reclassified on a scale of 1 to 10. For all datasets, except race, the higher values were given a score of 10 (or highest risk) and the lower values were given values of 1 (lowest risk). However, since race was based on ratio of black Americans who are at a lower risk, these values were reversed.

Combining the Data

These reclassified rasters were added together to create a new composite raster representing areas at risk for thyroid cancer in the United States. This involved first adding together gender and pregnancy values as pregnancy is directly dependent on gender. The data were then added together with the raster calculator using the formula:

Gender_reclass+ Preg_Reclass

The data was then reclassified with the values shown below:

<4	1
4-5	2
5-7	3
7-9	4
9-10	5
10-11	6
11-12	7
12-14	8
14-16	9
16-18	10

Since radiation testing was only relevant to people born before 1971, age needed to be taken into account in radiation calculations. For this analysis, raster calculator was used with the following formula:

90* Radiation + 10*Age

The age values used for this calculation were the ratio of the population over the age 33 in the year 2004, calculated with field calculator. The values of 90 and 10 were the weights given to radiation and age respectively. The value of 10 was used for Age because while one's risk for radiation in the United States is greater for people born before 1971 and thus should be taken into account, the nuclear testing in this time period is not the only source of radiation and thus should not be given too high of a weight.

This value was then reclassified based on Jenk's natural breaks and then added to the other risk factors to identify the areas most at risk using the following formula:

30*Gender_Preg + 15*Race + 15*Age + 40*Radiation_Age

The values of 30, 15, 15, and 40 in this calculation were the weights given to each risk factor based on how much a factor was predicted to affect one's risk of thyroid cancer (Mayo Clinic Staff, 2007).

The weight of 40 was chosen for radiation because of the known risk factors, unlike the other risk factors examined, it is known to actually "cause" papillary and follicular thyroid cancer and thus has the largest affect on one's chances of getting the disease.

The next highest weight of 30 was given to gender because although gender alone is not known to cause the disease, being female has been shown to significantly increase one's risk for these types of thyroid cancer.

Finally, race and age were given a weight of 15.These factors were given half the weight of gender because although there has been research linking one's risk for thyroid cancer to these

factors, the nature of this link and what affect it has on one's risk for the disease is largely unclear.

It should be noted that for the purposes of this analysis it was assumed that these factors represent one hundred percent of a person's risk for thyroid cancer; however it did not take into account other factors, specifically genetics, which could also play a role in one's risk for the disease.

A model for these calculations can be found in Figure 1 and the resulting map in Figure 2.

This data then was converted from a raster to a feature dataset using the Zonal Statistics tool in Spatial Analyst to identify the mean value for each state and then joining this table to the states dataset.

Results

This analysis provided a number of important pieces of information that could be useful in future studies. The map displaying the states at highest risk can be found in Figure 2.

Assuming that this model correctly identified the affect of certain risk factors on the disease, the map identified the areas with the highest at-risk populations. In this case, these areas are mainly focused in the Midwest and Northeast. This is in contrast to incidence values derived from the National Cancer Institute which had the highest values in the West (away from the ocean) and Northeast (Figure 3)

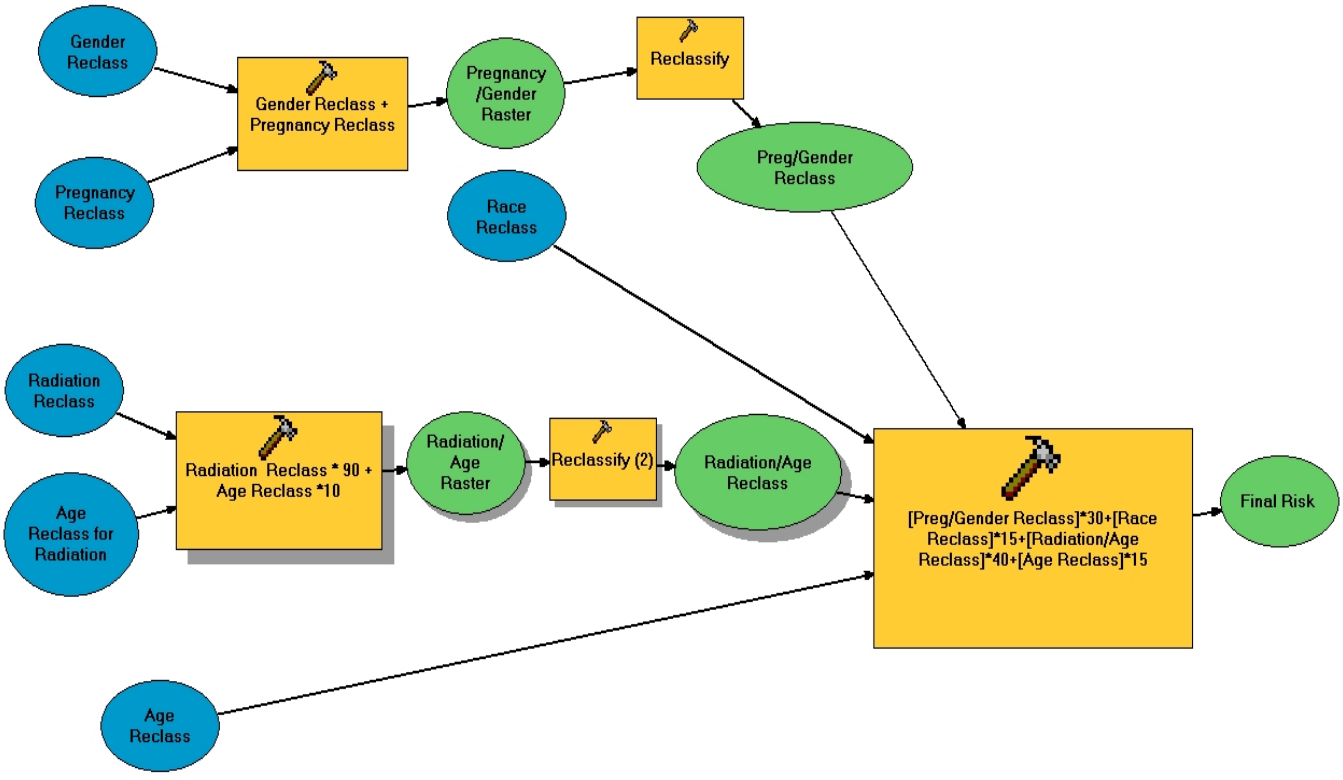

Figure 1. Model for risk analysis calculations.

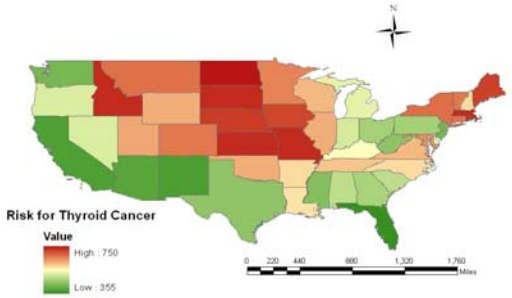

Figure 2. Map displaying states most at risk for thyroid cancer according to analysis criteria. Red areas have the most risk and green values have the least.

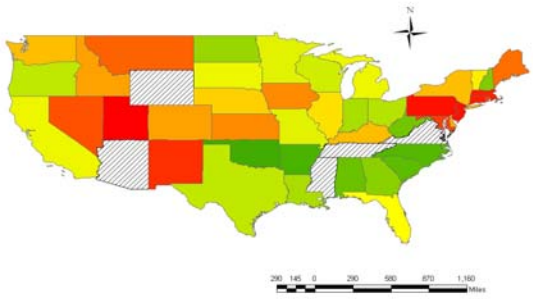

Figure 3. Map displaying 2004 incidence rates of thyroid cancer from the National Cancer Institute (2007a). The red values had the highest incidence rates while the green had the lowest incidence rates. The struck out values did not have data available.

To better understand the relationship between the risk factor map and the incidence rate map, another map was created for comparison. In order to compare incidence data that had incidence rates based on cases per 100,000 population per year and the risk factor data that was based on the sum of reclassified risk factors, the data sets had to be put on a similar scale. In order to achieve this, each state was ranked from 1 to 49 (including Washington D.C. and excluding Hawaii and Alaska) from highest actual and expected incidence rates to lowest actual and expected incidence rates. The difference between these two ranked datasets was then found by first joining the incidence and the risk factor layers. A new field was then created in the joined table and then the Field Calculator was used with the formula below to subtract the ranked fields for the two data sets.

RiskFactor_Rank - Incidence_Rank

From this calculation, a map was created to show what states had the greatest difference in rank between the calculated risk and incidence levels (Figure 4).

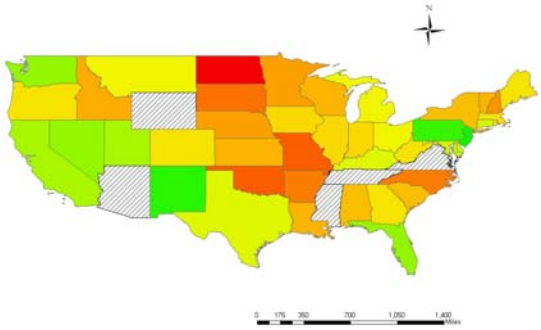

Figure 4. Map showing the difference between the ranking of states based on calculated risk and incidence levels for 2004. The red values represent a large difference with risk levels predicting a higher rank than the incidence levels would suggest. The green values represent a large difference with risk levels having lower ranks than incidence levels would suggest. Yellow represents ranks that are similar to what one would expect for that state. The struck out values are those where incidence values were unavailable.

Statistical Analysis

In order to test whether the differences in the risk factor and analysis ranks were statistically significant, the Spearman test was used.

The differences between the ranks

of incidence and ranks found through this analysis were used in the following formula to find the Spearman rank coefficient (Zar, 1999). With the hypothesis that H_0: $\rho_s \leq 0$ (there is no linear correlation between the datasets); H_A: $\rho_s > 0$ (there is a correlation between the datasets).

$$r_s = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

In this formula Σd_i^2 is the sum of the differences between these ranks squared and n is the number of samples taken for this test. In this analysis:

$$6\Sigma d_1^2 = 68590.5$$

$$n^3 - n = 44^3 - 44 = 85140$$

$$1 - (68590.5/85140) = .1944 = r_s$$

Since there were a number of tied ranks for x and y values, a correction for this tie had to be used. The following formula was used for this correction (Zar, 1999):

$$(r_s)_c = \frac{(n^3 - n)/6 - \sum d_i^2 - \sum t_x - \sum t_y}{\sqrt{[(n^3 - n)/6 - 2\sum t_x][(n^3 - n)/6 - 2\sum t_y]}}$$

Where n is the number of samples and t_x and t_y are calculated through the formulas below. Where t_i equals the number of tied values of X in a group of ties in the first formula and t_i equals the number of tied values of Y in a group of ties in the second.

$$\Sigma t_x = (\Sigma(t_i^3 - t_i))/12$$

$$((5^3-5)+(2^3-2)+(2^3-2)+(3^3-3)+(2^3-2)+(2^3-2)+(2^3-2)+(2^3-2))/12$$
$$= 15$$

$$\Sigma t_y = (\Sigma(t_i^3 - t_i))/12$$

$$((2^3-2)+(2^3-2)+(2^3-2)+(2^3-2)+(2^3-2)+(2^3-2)+(3^3-3)+(3^3-3)+(2^3-2)+(2^3-2))/12$$
$$= 8$$

From this, the calculated correction for the analysis becomes:

$$\frac{(44^3-44)/6 - 11431.75 - 15 - 8}{\sqrt{[(44^3-44)/6 - 2(15)][(44^3-44)/6 - 2(8)]}} =$$

$$\frac{2735.25}{14166.10} = .193 = (r_s)_c$$

From Table B.20 in Zar (1999):

$$(r_{s.})_{0.05(1),44} = .251$$

$$.25 > P > .10$$

Since the calculated value of .193 is less than the table value of .251 H_0 is not rejected and $\rho_s \geq 0$. This suggests that there is no significant correlation between the two data sets.

Discussion

Although the data from the risk factor analysis and incidence rates do not correlate, the information obtained in conducting this analysis can be used to better understand the distribution of thyroid cancer.

A risk factor analysis similar to the one used in this analysis allows researchers to identify areas where thyroid cancer rates might be higher than the incidence rates might suggest. The asymptomatic nature of this disease in its early stages means that it is often under diagnosed. By providing information and more accurate detecting technology to areas with these higher risk factors (in this case the Midwest and Northeast),

thyroid cancer may be diagnosed earlier and prevent problems later in life.

By pairing this risk factor analysis with the incidence data information, as in Figure 4, it is also possible to identify areas where the expected incidence and actual incidence do not match up. This makes it easier to identify other risk factors or give recognized risk factors a larger or smaller weight in the analysis. This is important with thyroid cancer as it may make it possible to identify sources of radiation not accounted for in a previous calculation. This study found the largest differences between incidence rate and risk factors in the West and Midwest. Future research should be focused in these areas so that a more accurate model and perhaps more accurate diagnoses could be made.

Conclusion

Understanding the nature of a thyroid cancer is one of the best ways to prevent it or at least lower morbidity and mortality resulting from this disease. GIS can be a useful tool in identifying possible risk areas so that resources and information can be given to areas that need them most. Moreover, by comparing risk areas with incidence rates, it is possible to more accurately determine the affect of particular risk factors on the incidence of this disease.

This study serves as a beginning in understanding thyroid cancer. It does not take into account genetics or other factors that are not distributed spatially. Moreover, the risk factors used, while based on published research, are not the only causes of thyroid cancer. This means that this study, especially due to its large scale, is not an indicator of whether or not a person will get thyroid cancer. Instead, it provides a basis for future studies so that more accurate detection and prevention actions can be taken so that fewer people will suffer the consequences of this disease.

Acknowledgements

Special thanks are given to the staff and students of the Resource Analysis program at Saint Mary's University in Winona as well the staff of Geospatial Services for providing me with the skills necessary to complete this project. I also would like to thank my family who supported me throughout this project. Finally, I would like to acknowledge the National Cancer Institute, who by making their information public domain, make projects like this possible.

References

American Cancer Society. 2005. *A Detailed Guide to Thyroid Cancer.* Retrieved June 22, 2007 from http://www.cancer.org/docroot/CRI/CRI_2_3x.asp?rnav=cridg&dt=43.

Davies, L. and Welch, H. G. 2006. Increasing Incidence of Thyroid Cancer in the United States, 1943-2002. *JAMA, 295,* 2164- 2167.

Mayo Clinic Staff. 2007. *Thyroid Cancer*. Retrieved June 20, 2007 from: http://mayoclinic.com/health/thryroidcancer/DS00492/DSECTION=2.

National Cancer Institute. 2002. *What you Need to Know About Thyroid Cancer.* Retrieved June 29, 2007 from http:// www.cancer.gov/cancertopics/wyntk/thyroid/page4.

National Cancer Institute. 2007a. *State Cancer Profile.* Retrieved November 4, 2007 from http://

statecancerprofiles.cancer.gov/cgi-bin/quickprofiles/profile.pl?00&080.
National Cancer Institute. 2007b. *Radioactive I-131 from Fallout.* Retrieved November 30, 2007, from http://www.cancer.gov/i131.
Slocum T.A., R.B. McMaster, F.C. Kessler, and H.H. Howard. 2004. Thematic Cartography and Geographic Visualizations, 2nd ed. Upper Saddle River, NJ: Prentice Hall, pp. 528.
Zar, Jerrold H. 1999. BioStatistical Analysis (Fourth Ed.). New Jersey: Prentice Hall, pp. 663.