DISCLAIMER OF QUALITY

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort to provide you with the best copy available. If you are dissatisfied with this product and find it unusable, please contact Document Services as soon as possible.

Thank you.

PAGE 2 IS MISSING FROM ORIGINAL THESIS
Analysis of community cancer mortality rates

by

Janice A. Vatland

S.B. Earth, Atmospheric, and Planetary Sciences
Minor Environmental Engineering Science
MIT, 1995

S.M.
MIT, 1996

SUBMITTED TO THE DIVISION OF BIOENGINEERING AND ENVIRONMENTAL HEALTH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Ph.D. IN TOXICOLOGY AND EPIDEMIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2001

© 2001 Massachusetts Institute of Technology, All rights reserved.

Signature of Author:  

Janice A. Vatland
Division of Bioengineering and Environmental Health
September 2001

Certified by:  

William G. Thilly
Thesis Supervisor

Accepted by:  

Ram Sasisekharan
Chairman, Committee on Graduate Students
Analysis of community cancer mortality rates

by

Janice A. Vatland

Submitted to the Division of Bioengineering and Environmental Health
September 2001
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

ABSTRACT

Residents of small communities or neighborhoods often observe what appears to them to be an excessively high number of cancer cases. Some of these observations are brought to the attention of public health officials in the hope that an environmental cause within these communities may be discovered and subsequently eliminated. However, what appears to be a cluster of cancer cases locally may be a chance occurrence when viewed from the perspective of the entire state or region. The primary aim of this work was to obtain and analyze a data set of sufficient magnitude to provide a means to discover if distributions of cancer mortality rates among communities within a particular state, for any of the most common forms of cancer, were compatible with chance.

To this end cancer mortality data for six of the largest states in the United States (California, Florida, Massachusetts, New York, Pennsylvania and Texas) were collected, converted into mortality rates and analyzed to discover if the variations among communities could be accounted for by chance alone. These data comprised one third of all recorded deaths in the period of approximately 1969 to 1998. The 21 most common forms of adult and 6 most common forms of pediatric cancers were organized to permit analyses within each of the 6 states with regard to age (0-19, 65-84 and ≥ 85 years), gender and ethnicity (European American and Non-European American descent).

Key to this work was the mode of statistical analysis. For each community an expected mortality rate and its expected distribution was defined by the average mortality rate for all communities within each state and the binomial distribution, respectively. These expected distributions were summed for all communities to define the expected chance distribution of community mortality rates for each state, cancer, gender, age cohort and ethnicity. This produced nearly 800 separate “chance” distributions. Each of these was compared to the corresponding observed distribution using the Kolmogorov-Smirnov statistical test. This test was designed to discover statistically significant differences between any two distributions. Here it was used to determine which of the nearly 800 observed distributions could not be accounted for by chance alone. Of these
comparisons, 16 were found to have observed distributions significantly different from
the expected by chance distributions. All 16 had distributions that exhibited greater
dispersion than expected by chance, and none had distributions with less dispersion than
the chance expectation.

The primary conclusion of this work is that for nearly all adult and pediatric cancers, the
variation in mortality rates among communities in each of the 6 states analyzed are
wholly consistent with what would have been expected by chance alone. This implies
that variation among environmental and genetic risk factors among communities had no
significant effect on a wide variety of cancer types over the past approximately thirty
years of observations. This is true for each of the three age groups, two genders and two
ethnic groups analyzed.

With regard to the 16 exceptions where the Kolmogorov-Smirnov analyses indicated a
greater dispersion of cancer mortality rates than expected by chance alone, 12 were
distributions of lung cancer mortality rates in adults aged 65-84 years for both genders
and ethnic groups. This outcome was not unexpected because of the previously
documented “urban effect” wherein lung cancer mortality rates were higher in urban than
in suburban and rural communities. This difference has been attributed in a large part to
higher prevalence of cigarette use in urban areas. This also indicates the sensitivity of the
Kolmogorov-Smirnov method to detect the presence of risk factors that vary throughout
the communities. The 4 other examples where significant differences between the
observed and expected chance distributions were found for: prostate cancer in TX for
NEAM (Non-European American males), prostate cancer in CA for EAM (European
American males) and NEAM and breast cancer in CA for EAF (European American
females), all aged 65-84 years. The reason for these differences is not yet known.

A second independent means to identify communities with significantly different cancer
mortality experiences was developed. The cancer mortality rates for all 2216
communities in the set for all 6 states were explored for five major forms of adult and
two forms of pediatric cancer for European American males and females for the age
groups 65-84 years and 0-19 years, respectively. The binomial distribution was utilized
to calculate the probabilities for the observed number of deaths for each cancer and
cohort given the community’s cohort population size and the state average mortality rate.
To correct for the number of comparisons, the probability level (for observing a
significant difference by chance for each comparison) was set at 1.13 x 10^-5 according to
the Bonferroni limit for the two-tailed 95% confidence interval, or 0.025 / 2216. Using
this method, of the nearly 27,000 trials for the communities, cancers and cohorts, 566
trials were found to be significantly different. Of these significant trials, nearly 55% had
already been identified from the Kolmogorov-Smirnov analyses of the community cancer
mortality rate distributions. Of the remaining 257 significant trials, 91 occurred for
cancer/cohort groups that had exhibited a shift in the observed and expected by chance
distributions which indicated an urban/rural difference. Maps of the significant
communities also indicated that significant differences occurred due to residence in urban
locales. Therefore, outside of the “urban effect”, the results of the individual community
analyses also indicate community cancer mortality rates are distributed by chance.
Any data and mathematical and/or computer tools mentioned herein can be acquired by contacting Janice A. Vatland, at kroll@alum.mit.edu, or Professor William G. Thilly, at thilly@mit.edu. The offices of the MIT Center for Environmental Health Sciences are located in Building 16, Room 743 and can be reached at the following phone number: 617-253-6220.

Thesis Supervisor: William G. Thilly
Title: Professor of Toxicology
ACKNOWLEDGMENTS

This thesis represents the culmination of many hours of work. The compilation, organization and analysis of the most comprehensive data set known to date of community cancer mortality records cannot be credited to me solely as this enterprise was facilitated by many persons whom must be recognized. Many thanks, of course, to the individuals who helped provide the mortality and population data necessary for analysis. The following individuals provided the individual computerized death records for the six states analyzed: Elaine Trudeau and Charlene Zion from the Massachusetts Vital Registry, Mary Anne Freedman and Robert Krasowski both from the National Center of Health Statistics as well as Rod Palmieri from the Department of Health and Human Services in California, Gary Sammet from the Florida Department of Health, Craig Edelman from the Pennsylvania Department of Health, Gene Therriault and Robert Draiss from the State of New York Department of Health as well as Louise Berenson from the New York City Department of Health, and Leland Carmichael from the Texas Department of Health. Thanks must also be extended to those who helped to obtain the corresponding population data. Special gratitude for providing the population data must go to Marie Pees from the United States Census Bureau Department of Special Tabulations. Ms. Pees spent many hours extracting the necessary population files for the communities within the six states for which mortality data had been obtained. The Boston Regional Office of the Bureau of the Census was also extremely helpful in obtaining supplementary population data.

Gratitude must also be extended to my thesis committee. All of the professors in addition to my advisor, Professor William Thilly, provided me with many helpful suggestions as well as a high level of support and encouragement. Professor Stephan Morgenthaler deserves special mention as he traveled extensively to be present at the various committee meetings in addition to the thesis defense. Professor Morgenthaler also provided much statistical guidance throughout the various stages of data analysis. The greatest amount of appreciation, however, must be reserved for my advisor, Professor Thilly. It has been a tremendous honor to be under his tutelage. He provided me with not only a tremendous opportunity but also taught me many important lessons. The most
important of these is to always be critical of widely held assumptions. This idea was not only applied to this thesis work but has also defined my approach to other widely held but unproven dogma.

Thanks must also be extended to all of the past and present members of the Thilly lab with whom it has been a great pleasure to work with. These colleagues not only made the many research hours pass more quickly and enjoyably but also were a great source of support and advice. There are many individuals that should be listed here but for fear of inadvertently excluding someone a complete list will not be attempted. I would, however, like to mention those colleagues who were extremely supportive especially during the last several grueling months. Dr. Luisa Marcelino, Dr. Brindha Muniappan and Andrea Kim were especially wonderful and provided the necessary encouragement at critical times. Special thanks must also go to my colleague and longtime friend Dr. Pablo Herrero-Jimenez. Pablo has been a good friend from as early as my first undergraduate day at M.I.T. It was tremendously special not only to share our undergraduate careers at M.I.T. but also our doctoral careers within the Department of Bioengineering and Environmental Health at M.I.T.

Finally, my family must be duly recognized. It goes without saying that none of this could have even been attempted without their support and encouragement. Many thanks to my sister Kerrie for always making me smile even when it was difficult to do so. Thank you also to my father, Gerald, for always believing in me and making me feel wonderful about whatever I accomplished. It is, however, my mother, Astor, who must be credited as the impetus for all of this work. Without her high expectations none of this would have even been attempted or considered. The importance she placed on education inspired me at an early age to always strive for the best. Lastly, much gratitude must also be reserved for my husband, Chad Kroll, who suffered for many years as I journeyed through my educational career. Without his help and care it would not have been possible to survive the many trials and tribulations that were endured in the pursuit of this degree.
TABLE OF CONTENTS

Section                                  Page
Title Page                                1
Committee Page                            2
Abstract                                  3
Acknowledgments                           6
Table of Contents                         8
List of Figures                           10
List of Abbreviations                     13
List of Symbols                           14
1. Introduction                           15
   1.1 Love Canal                          26
   1.2 Woburn, Massachusetts              27
   1.3 Los Alamos, New Mexico             28
   1.4 Cape Cod, Massachusetts            29
2. Mathematical Basis of Analyses         31
   2.1 Definition of the Null Hypothesis   31
   2.2 Definition of the Chance Expectation 31
   2.3 Definition of the Observed Distributions 41
   2.4 Comparison of the Chance Expectation and Observed Distributions 41
       2.4.1 Calculation of $D^*$, Application of the K-S Test for Differences Between Distributions 41
       2.4.2 Comparison of $D^*$ and the Critical Value $\varepsilon^*$ Distributions 44
3. Historical and Geographical Variations 48
   3.1 Historical Variations               48
   3.2 Geographical Variations             58
4. Data                                   67
   4.1 Acquisition, Compilation and Organization 67
   4.2 Accuracy                             80
5. Analyses of Data                       84
   5.1 National versus Six-State Mortality Rates 84
   5.2 Distribution Analyses               98
       5.2.1 Consideration of Weighted and Unweighted Mortality Rate Averages 98
       5.2.2 Distribution Analyses Results 106
       5.2.3 Histogram Interval Size 130
       5.2.4 Accuracy and Detection of Perturbations of the K-S Method 133
       5.2.5 Analyses for Individual Communities 141
6. Discussion                             155
   6.1 Distribution Analyses               155
   6.2 Statistically Significant Individual Communities 157
   6.3 Space-Time Considerations, Breast Cancer as an Example 163
       6.3.1 Consideration of Spatial Clustering 163
       6.3.2 Consideration of Temporal Clustering 167

8
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States age-specific mortality rates for malignant neoplasms as a function of birth cohort for European American males (EAM) and European American females (EAF).</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>United States age-specific mortality rates for lung cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>United States age-specific mortality rates for stomach cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>United States age-specific mortality rates for breast cancer as a function of birth cohort for European American females (EAF).</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>Idealized observed and expected cancer mortality rate distributions depicting the expected due to chance variation and the observed if risk factors vary among the set of communities.</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>Ratio of the observed and expected number of cancers deaths for breast cancer in Massachusetts' 251 cities and towns versus the log of the cohort population size for European American females aged 65 to 84 years in the period 1969-1995.</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>Illustration of the cumulative distributions of the observed and expected by chance and the determination of the D* Statistic for the Kolmogorov-Smirnov (K-S) test.</td>
<td>42</td>
</tr>
<tr>
<td>8</td>
<td>United States age-specific mortality rates for lung cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>United States age-specific mortality rates for stomach cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>United States age-specific mortality rates for breast cancer as a function of birth cohort for European American females (EAF).</td>
<td>54</td>
</tr>
<tr>
<td>11</td>
<td>United States age-specific mortality rates for prostate cancer as a function of birth cohort for European American males (EAM).</td>
<td>56</td>
</tr>
</tbody>
</table>


14: Age-specific smoking prevalence comparison between urban and rural communities in Massachusetts.

15: A comparison of 1-Survival versus mortality rate values for prostate cancer for 82 EAM, 60 to 64 (62.5) years of age in the United States.

16: Composite age-specific mortality rates for the six states and the United States for EAM, EAF, NEAM and NEAM for seven forms of cancer.

17: Observed and expected by chance stomach cancer mortality rate distributions for the 477 New York cities/towns, EAM, 65-84 years of age.

18: Observed and revised expected by chance stomach cancer mortality rate distributions for the 477 New York cities/towns, EAM, 65-84 years of age.

19: Observed and expected by chance distributions for the 21 adult and 6 pediatric forms of cancer for EAM (with the exception of the female specific cancers) for the state of Pennsylvania.

20: Values for the D* statistic for the Kolmogorov-Smirnov test as the number of histogram intervals were varied for three cancers and cohorts.

21: Perturbations of the Massachusetts pancreatic data set for EAM, 65-84 year olds and the level of detection (D*) with the K-S method.

22: Ratio of the observed and expected number of cancer deaths for lung, colon and breast cancer for all 2216 cities and towns versus the log of the cohort population size for European American males aged 65 to 84 years for all available death years (European American females were considered for breast cancer).

23: Communities with at least one significant trial for seven major adult and two major pediatric cancers for EAM and EAF.

24: Communities with at least one significant trial for six major adult and two major pediatric cancers for EAM and EAF. (Lung cancer results were excluded)

25: Spatial map of community specific mortality rates for breast cancer.

26: Spatial map of communities with one or more significant time intervals for breast cancer among EAF.


29: Massachusetts age-specific lung cancer mortality rates for EAF, death years 1969-1995 according to population density.

30: Distribution of colon and breast cancer mortality rates for EAM and EAF, respectively for the 50 States and the District of Columbia for the death years 1968-1991.

Program 1: Program to calculate the theoretical distribution over a range of observed cancer rates, obs, as defined by the binomial distribution and the cohort population values, pop, and average cancer rate, p.

Program 2: Example of a Matlab™ program used for mortality data matrix creation.

Program 3: Example of a Matlab™ program used for mortality data extraction from matrices.

Program 4: Example of a Matlab™ program used for data extraction for the population files.

Program 5: Example of a Matlab™ program used for population data extraction by desired age group.

Program 6: Mathematica™ program for the random simulation of the number of deaths assuming a specified binomial distribution.

Program 7: Simulation program to test the difference between the simulated probability level and the Bonferroni limit.

Table 1: D* test statistic values for all comparisons performed for each state.

Table 2: Number of communities for which the observed mortality rates had a probability of less than that defined by the 95% confidence level in conjunction with the Bonferroni limit (p<2.6 x 10^-5) for the listed cancers.
LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRFSS</td>
<td>Behavioral Risk Factor Surveillance System</td>
</tr>
<tr>
<td>CA</td>
<td>California</td>
</tr>
<tr>
<td>CD-ROM</td>
<td>Compact Disk-Read Only Memory</td>
</tr>
<tr>
<td>CNS</td>
<td>Central Nervous System</td>
</tr>
<tr>
<td>EA</td>
<td>European American</td>
</tr>
<tr>
<td>EAF</td>
<td>European American females</td>
</tr>
<tr>
<td>EAM</td>
<td>European American males</td>
</tr>
<tr>
<td>FL</td>
<td>Florida</td>
</tr>
<tr>
<td>ftp</td>
<td>International Classification of Disease</td>
</tr>
<tr>
<td>JF</td>
<td>Japanese female</td>
</tr>
<tr>
<td>JM</td>
<td>Japanese male</td>
</tr>
<tr>
<td>NHL</td>
<td>Non-Hodgkin's Lymphoma</td>
</tr>
<tr>
<td>K-S</td>
<td>Kolmogorov-Smirnov</td>
</tr>
<tr>
<td>MA</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>MA DPH</td>
<td>Massachusetts Department of Public Health</td>
</tr>
<tr>
<td>MITVMA</td>
<td>National Center of Health Statistics</td>
</tr>
<tr>
<td>NCHS</td>
<td>Non-European American</td>
</tr>
<tr>
<td>NEAM</td>
<td>Non-European American males</td>
</tr>
<tr>
<td>NEAF</td>
<td>Non-European American females</td>
</tr>
<tr>
<td>NY</td>
<td>New York</td>
</tr>
<tr>
<td>PA</td>
<td>Pennsylvania</td>
</tr>
<tr>
<td>SEER</td>
<td>Surveillance, Epidemiology, and End Results</td>
</tr>
<tr>
<td>TCE</td>
<td>tri-chloro ethylene</td>
</tr>
<tr>
<td>TX</td>
<td>Texas</td>
</tr>
<tr>
<td>USGS</td>
<td>United States Geological Survey</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS

\( \alpha \)  probability of observing a chance occurrence
\( \gamma \)  probability of observing at least one significant trial
1-S  1-probability of survival
d  distance between cities
D*  maximum difference between the cumulative observed and chance distributions
D  difference between the cumulative observed and chance distributions for any histogram interval
death(a, y)  number of persons who died from all causes age a at time y
\( \varepsilon^* \)  critical value for Kolmogorov-Smirnov test
\( F_1(x) \)  cumulative observed distribution
\( F_2(x) \)  cumulative chance distribution
intpop(a, y)  interpolated population estimate for a cohort of age a+1 at time y+1
lat1  latitude of city 1
lat2  latitude of city 2
long1  longitude of city 1
long2  longitude of city 2
n  cohort population size, Equation 1
n  number of trials (communities), Equations 5 and 6
N  total number of communities
Obs(h, t)  observed mortality rate for those born in a time period, h, who died during the age period, t
p  cancer rate, Equation 1
p  confidence level of interest, Equation 4 and 5
\( P_Y(y) \)  probability of observing the number y, Equation 1
pop(a, y)  population for a cohort of age a at time y
1. Introduction

Epidemiology, the study of the frequency and distribution of disease in the human population, is an important area of research that seeks to identify groups at risk for disease and to understand the influence of inherited and environmental factors on such risk. To better understand the etiology of cancer, researchers have long used epidemiological studies of cancer mortality and/or incidence rates to provide hypotheses about the risk factors for carcinogenesis. These studies have successfully uncovered risk factors such as smoking (lung cancer) [1-2], sunlight (skin cancer) [3-5], asbestos (mesothelioma) [6-7] and vinyl chloride (liver angiosarcoma) [8-9]. The influence of these factors on cancer risk was discovered with epidemiological studies through cohort and case-control analyses. For instance, the link between smoking and lung cancer was discovered not only through the recognition of a concomitant rise in lung cancer mortality in birth cohorts following the introduction of cigarettes but also to case-control studies which showed a high fraction of smokers among lung cancer victims [1-2]. Risk factors, such as vinyl chloride, were discovered through studies comparing exposed workers with those unexposed. Vinyl chloride was shown to increase the mortality risk from the rare cancer, liver angiosarcoma, through the analysis of standard mortality ratios in workers in the vinyl chloride industry [8-9]. Additionally, vinyl chloride was shown to exhibit a dose response for liver angiosarcoma in nested case-control studies of these workers [9]. However, although these studies were successful in providing evidence for risk factors involved in carcinogenesis, these risk factors, exclusive of lung cancer, are responsible for little of the cancers that occur in the general population. Death from skin cancer represents an extremely small percentage of deaths from all forms of malignant neoplasms [10] and exposure to the other chemical risk factors are confined to the workplace.

Epidemiological studies have also uncovered strong associations between cancer and other risk factors such as lifestyle habits. It has been shown that an increase in overall cancer risk is associated with increased obesity and inactivity [11-12]. This association does not, however, pinpoint the real risk factor as obesity and inactivity not only have an
implied correlation but may also be confounded by other possible risk factors such as dietary habits including overall caloric intake. Breast cancer has been shown to be associated with fertility rates [13] and delayed childbearing [14-15]. Although there have been many studies to support these associations, the exact nature of the risk factors is not well-defined. Certainly the mechanisms by which these factors or their possible confounders may be acting are unknown. The epidemiological studies that have successfully defined risk factors involved in carcinogenesis in the general population have been confined to risk factors that result in a high relative risk in the exposed and low rates of occurrence for the general population. In all, for the large fraction of the population that will die of lung cancer the exact nature of the carcinogenic process is unfortunately unknown.

In addition to this lack of knowledge regarding the etiology of cancer in the general population, the increasing risk of death due to cancer over the last century has served to hold the interest of scientists and the public alike. Malignant neoplasms at the beginning of the 20th century accounted for about 6% of all deaths and now account for over 24% of all deaths [10-16]. This increasing risk for mortality due to malignant cancers is best illustrated with the birth year and age-specific cancer mortality rate curves shown in Figure 1 for European American males (EAM) and European American females (EAF), respectively. Figure 1 shows the rise in age-specific mortality rates for all malignant cancers beginning with the birth cohort in the 1800s and ending with the birth cohort of the 1970s. The rise is more dramatic in males but also occurs in the female population. These mortality data were compiled and organized by Dr. Pablo Herrero-Jimenez utilizing data from the United States Department of Health and Human Services and United States Census Bureau [10-16]. These data as well as the historical data for many of the major forms of death can be accessed on the web site http://cehs4.mit.edu.

Risk of mortality has been changing for many other forms of cancer. Most of these changes have been increases in age-specific mortality rates as a function of history, although there are some examples where the age-specific mortality rates have been decreasing or have remained constant. Lung cancer is the most obvious example of a
form of cancer that has increased dramatically as a function of history. Stomach cancer is an example of a form of cancer that has exhibited marked decreasing risk for cancer mortality. Breast cancer is one example where age-specific mortality rates have remained constant over the past 150 years. These changes are illustrated in Figures 2 to 4. Figure 2 shows the age-specific mortality rates for lung cancer which clearly show dramatic increases as a function of history, and Figure 3 shows the decrease occurring for stomach cancer in the United States. The last figure, Figure 4, for breast cancer shows the fairly stable mortality from this form of death. Although there are other examples of cancers that are decreasing or remaining stable as a function of history the majority of more common cancers have been increasing. These include (in addition to lung cancer): leukemia, lymphoma, central nervous system (CNS) cancer, pancreatic cancer, kidney cancer, ovarian cancer and prostate cancer.
Figure 1: United States age-specific mortality rates for malignant neoplasms as a function of birth cohort for European American males (EAM) and European American females (EAF).

Obs (h, t) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
Figure 2: United States age-specific mortality rates for lung cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).

Obs (h, t) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
Figure 3: United States age-specific mortality rates for stomach cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).

Obs \((h, t)\) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, \(h\), who died during the age period, \(t\).
Figure 4: United States age-specific mortality rates for breast cancer as a function of birth cohort for European American females (EAF).

Obs (h, t) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
To further understand the etiology of carcinogenesis in the general population researchers have used community studies in the hope risk factors, particularly environmental, may be identified. This has coincided with the increased interest on the part of the public due to the recognition of the increasing proportion of the population dying of cancer and the attention paid to local environmental pollution. There have been numerous requests for the investigation of cancer rates in cities and towns across the country. In 1989, there had been between 1300 and 1650 formal requests to investigate possible cancer "clusters" nationwide [17]. More recently, in 1997, there were 1100 formal investigation requests made mostly by concerned citizens in a survey of state public health departments [18]. In Massachusetts alone, the public health department in 1998 responded to between 3000 and 4000 calls made by alarmed and questioning residents [19]. The result of community residents' awareness of local pollution and sensitivity to the number of disease cases that occur within their communities has been the investigation of many cities, towns and neighborhoods across the country seeking to identify environmental causes for disease. Examples of the most famous of these cases within the United States are Love Canal, the Woburn leukemia cluster, Los Alamos and Cape Cod.

1.1 Love Canal

For decades Love Canal, a part of Niagara Falls, New York, was a waste site for Hooker chemical company. The waste from this site, known to be a variety of caustics, alkalis and chlorinated hydrocarbons from the manufacture of dyes and solvents, had been seeping into the basements of the residents' homes and was visibly polluting the residents' yards as well as the community's school and playgrounds [20-21]. In the 1970s, the resulting controversy and public outcry from the obvious pollution and reports suggesting excess miscarriage rates [22] and chromosomal aberrations [23] for some Love Canal residents resulted in the relocation of the majority of residents as well as the destruction of the 100 homes nearest to the waste site [20]. Although the reports suggested risk to Love Canal residents, a study comparing cancer incidence from 1955 to 1977 for the Love Canal census tract with the rest of New York state, exclusive of New York City, found there was no excess increases for cancers overall for Love Canal [24].
Additionally, even though lung cancer seemed to be increased for the census tract, they found that this increase did not correlate with proximity to the waste site [24]. Even without confirmed evidence of risk to the Love Canal residents, the controversy resulted in the allocation of approximately 20 million dollars and affected the lives of a thousand families [20]. Unfortunately the claims of health effects are generally remembered as the discovery of health effects despite the evidence contrary to those claims.

1.2 Woburn, Massachusetts

The Woburn leukemia cluster, ultimately the topic of numerous scientific studies as well as a number of books and even a motion picture, began with an investigation by residents led by Anne Anderson [25-26]. Like Love Canal, the Woburn community was adjacent to industrial waste sites. Additionally, the residents had noticed a suspicious number of childhood leukemia cases in their community. Responding to community concern, a number of studies were conducted by the Massachusetts Department of Public Health (MA DPH). The MA DPH concluded that 12 leukemia cases, from 1969-1979, in Woburn represented a statistically significant increase by comparing the observed number of cases relative to the expected derived from the Third National Cancer Survey [27] assuming a Poisson distribution [28]. They argued for the number of cases among pediatric males in Woburn the observed number was 5.2 times higher than the expected and calculated that such an occurrence would have occurred with a chance probability of less than $7 \times 10^{-3}$ [28]. This estimate, however, did not account for the large number of communities in Massachusetts.

The authors suggested that the occurrence of leukemia, particularly in the eastern part of the town, was possibly due to exposure to water from the municipal wells, Wells G and H [28]. With the link to the well water, hypotheses were formed regarding specific pollutants found in the water, such as the metals, arsenic and chromium, as well as the solvent tri-chloro ethylene (TCE). Research was conducted to explore the link between these pollutants and the leukemia cases. However, the subsequent research failed to find any link between the purported arsenic and chromium levels in hair analyses and access
to water from Wells G and H [29]. Another study claimed that only half of the cases showed a positive association to the well water from Wells G and H as half of those with leukemia were residents of the western part of Woburn who received their drinking water from another source [30]. These investigations provided no evidence to support the link between the exposure to water from Wells G and H and the increase in leukemia rates in Woburn. Furthermore, no study yet has demonstrated any evidence that the leukemia cases were caused by environmental factors.

1.3 Los Alamos, New Mexico

In Los Alamos, New Mexico a seemingly high number of brain cancer cases were originally reported by a community resident in 1991 [31]. Due to the public's interest, the New Mexico Department of Health investigated the brain cancer incidence rates over two decades for this area and found that the frequency of brain cancer for that area, though higher than the state average, fell well within the range expected by chance around all New Mexico communities [32]. Kulldorff, et. al., 1998 analyzed these findings and also concluded that the "cluster" fell within chance [31]. Using incidence data for malignant neoplasms of the brain and central nervous system and population data for New Mexico counties, Kulldorff et. al., 1998 determined that the Poisson probability values for individual and groups of counties throughout various time points within the period of 1973 through 1991 were not significant even for the areas that were the most likely clusters (areas with the lowest Poisson probabilities). A space-time scan statistic was used for this study and probability values were compared with values created with Monte Carlo simulations. Most importantly the space-time scan statistic took into account the problem of multiple testing as many counties, combination of counties and time periods exist for study. These multiple trials would cause many areas to appear to be significantly different from the expected; however, many of these would have occurred by chance outside of standard 95% confidence intervals.
1.4 Cape Cod, Massachusetts

Finally, in Massachusetts, studies now are underway investigating the purportedly elevated incidence rates of breast cancer on Cape Cod. Cape Cod breast cancer incidence rates have been reported to be significantly higher on Cape Cod for the period 1982-1990 than the rest of the state with data from the MA DPH [33]. Of the eight towns found to be at least 25% higher than the state average (p<0.001) seven occurred on Cape Cod. An additional two towns were found to be significantly different from the state average with the significance criterion p<0.05. Due to preliminary studies as well as the intense public interest, the Silent Spring Institute was founded in 1994 and allocated initially $1.5 million to find the environmental cause for the increase in breast cancer on Cape Cod [34]. The Silent Spring institute is now in the process of screening hundreds of possible environmental pollutants that may be at different levels on Cape Cod than the rest of the state. However, these studies did not account for the distribution of cases throughout Massachusetts nor did they independently explore breast cancer mortality rates.

Although there have been many more studies of individual communities than discussed above, none of these studies have led to the discovery of any carcinogenic risk factor. Thus, what may have appeared to be a significant local increase in an individual community, may have simply been a chance occurrence. Studies that have utilized larger data sets and that have accounted for the problem with multiple comparisons have found that areas that appear to have significantly higher cancer rates at a local level are simply due to chance in the whole system. Kulldorff’s et. al., 1998 study of brain and central nervous system cancer illustrated this with a space-time scan statistic. This study is relatively recent; however, the idea to utilize sufficiently large data sets and to account for the number of comparisons for any set of communities is not. Leukemia cases in Connecticut among the 169 towns in that state for the period of 1945 to 1959 [35] were studied. The study was performed by calculating and comparing the probabilities in space and time, in this case at the town level for the number of leukemia incidences registered each year for children under the age of 15. It was concluded that no community in Connecticut had a leukemia experience outside of chance expectation.
Likewise, studies in Sweden have shown that the cancers around nuclear sites occur randomly [36]. Other studies have also demonstrated that for both pediatric leukemia and astrocytoma, a form of brain cancer, geographic “clustering” of areas with elevated rates occurs by chance [37-38]. In addition to space-time statistics, the distributions for a set of community cancer rates may also be analyzed. One study, which used the observed and expected Poisson distributions for 100 small researcher-defined “communities” within four counties in California, found that the distribution of childhood cancers were not significantly different from chance expectation [39]. The comparison of the distributions over the set of communities for the period of 1980-1988 were not significantly different using the Chi-square test. Although these studies appropriately sought to define community cancer risk using data for a number of communities and accounted for the multiple comparisons, these studies were still relatively limited in scope. In order to more completely characterize general cancer risk at the community level in the population it is necessary to explore many forms of cancer for larger numbers of communities.

Therefore, in order to determine community cancer risk and to identify communities that are candidates for additional analysis, a large community data set is warranted with a method that can determine whether or not the data are distributed by chance. If community cancer rates are found not to be distributed by chance variation alone, this would be prima facie evidence that cancer rates among communities are affected by risk factors and signal the need for additional research. Additionally, it is also important to scrutinize the probability an individual community is significant by chance prior to performing studies within that community. The use of statistical limits must be complemented with methods to adjust for multiple comparisons to decrease the number of communities deemed significant by chance. Finally, it is also important to appreciate that although significant differences may be found among communities’ cancer rates, it would be premature to assume these differences are necessarily environmental in nature. There are many factors that could vary among communities, which include genetic factors (family groups), medical care/socioeconomic differences as well as environmental factors.
2. Mathematical Basis of Analyses

2.1 Definition of the Null Hypothesis

The analysis of the distribution for a set of communities requires a straightforward method to discover whether or not community cancer rates are distributed as expected by chance alone. This could be an efficient method to discover evidence of significantly elevated community cancer risk. Significant variation of risk factors among a set of communities would alter the observed distribution of cancer rates in such a way that would distinguish it from the distribution expected due to chance variation alone. This is illustrated in the idealized graph, Figure 5. The variation of risk factors across a set of communities would cause the observed distribution to be broader with a lower maximum than what would be expected by chance alone. Therefore, to test the null hypothesis (that community mortality rates are distributed due solely to chance variation), the observed cancer rate distributions for a large set of communities can be simply compared to the chance expectation. To perform this analysis it is necessary to develop a method to calculate the expected cancer rate distributions for any cancer and cohort for a set of communities as well as a method to compare the observed and expected distributions.

2.2 Definition of the Chance Expectation

The expected number of cancers that occur in small populations will naturally exhibit large chance variations. In order for a significant difference in the community’s cancer experience to be found requires an increasingly greater fold difference in the observed number of cancers relative to the expected as the population size decreases.

For a population of 100,000 persons and a probability of 100 cancer cases per 100,000 persons a 30% increase in the number of cases would be significant according to the binomial distribution given in Equation 1 below (using the 95% confidence limits). However, for a population size of 10,000 there must be at least an 80% increase in the number of cases to be deemed significant. As the population size decreases further the
Figure 5: Idealized observed and expected cancer mortality rate distributions depicting the expected due to chance variation and the observed if risk factors vary among the set of communities.
The mortality rate is defined as:

\[
\text{Mortality Rate} = \frac{\text{Recorded number of deaths in a community for a cancer type and cohort}}{\text{Total persons alive in the community for the cohort}}
\]
\[ P_Y(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \]  

Equation 1

where,

- \( P_Y(y) \) = the probability of observing \( y \) cases of cancer
- \( n \) = is the size of the cohort population (the estimated population for the time period and cohort of interest)
- \( p \) = the probability an individual is afflicted (the average cancer rate)

increase in the ratio of the observed number of cases versus the expected required to be statistically different also widens. Therefore, for population sizes such as small towns or neighborhoods the fold-increase that may be observed by chance can be quite large. It would be common to observe a several-fold increase in small towns and/or neighborhoods relative to a state average. Figure 6 shows the data for a set of 251 communities in Massachusetts which illustrates the wider variation in the values of the observed number of cancer deaths relative to the expected as the cohort population size decreases. The data depicted are for breast cancer for those aged 65-84 years who died in the period from 1969 to 1995. Although the data were for cities and towns, smaller areas such as neighborhoods would be expected to exhibit even larger variations in the observed, relative to the expected number of cancer deaths based on the state average.

Analyses of the distributions of cancer rates among large sets of communities is an approach that can determine whether or not cancer rates are distributed by chance variation alone. Such analyses would provide *prima facie* evidence for risk factors at the community level if they occurred. In order to determine whether or not the observed distributions of community cancer rates were distributed according to chance variation alone it was important to define the null hypothesis, \( H_0 \). The null hypothesis as well as the method of defining the expected by chance distribution to be compared with the observed is detailed on the next page.
Figure 6: Ratio of the observed and expected number of cancers deaths for breast cancer in Massachusetts’ 251 cities and towns versus the log of the cohort population size for European American females aged 65 to 84 years in the period 1969-1995.

As the log of the cohort population size decreases the ratio of the observed number of cancers and the expected increases in magnitude.
$H_0$: the observed distributions of cancer mortality rates for a set of communities within a state and a fixed age group, ethnicity, gender and type of cancer are not significantly different from what is expected by chance alone.

For a fixed age group, ethnicity, gender and type of cancer the following definitions for a set of communities were as follows:

Communities: 1, 2, ..., k
   \[ i = 1, 2, ..., k \]

Number of Deaths for each Community:
   \[ d_1, d_2, ..., d_k \] (summed over a specified period of time)
   \[ d = d_1 + d_2 + ... + d_k \]

Number at Risk for each Community (number alive in the period):
   \[ n_1, n_2, ..., n_k \] (summed over a specified period of time)
   \[ n = n_1 + n_2 + ... + n_k \]

Mortality Rate:
   \[ d_1/n_1, d_2/n_2, ..., d_k/n_k \]

Distribution of the Mortality Rates Across Communities:
   \[ r_0 < r_1 < r_2 < ... < r_j \]

For a fixed community, $i$, the theoretically expected distribution is defined by:
   \[ P(r_j < \text{Mortality Rate of Community, } i < r_{j+1}) = P(r_j < D_i/n_i < r_{j+1}) \]
   \[ = P(r_j n_i < D_i < r_{j+1} n_i) \]
where, $D_i \sim B(n_i, \rho)$ where, $\rho$ is the average rate acting in a given state

Ultimately the fraction of communities with mortality rates in the range $r_j < r < r_{j+1}$ is equal to:

$$
\frac{1}{k} \sum_{i=1}^{k} P(r_j n_i < D_i < r_{j+1} n_i) \quad \text{where} \quad D_i \sim B(n_i, \rho) \quad \text{Equation (2)}
$$

Note: Since the state mortality rate is a weighted average of the community rates, an alternate definition of the theoretical distribution could have been based on an unweighted average of the community mortality rates. This will be described further in Section 5.

The theoretical distributions were calculated with a Mathematica™ program using data from each set of communities. An example of this program is given as Program 1.
Program 1: Matlab™ program to calculate the theoretical distribution over a range of observed cancer rates as defined by the binomial distribution, the cohort population values and average cancer rate.
<<Statistics`DiscreteDistributions`

pop = {107733,30917,129564,258222,52669,101446,22010,15201,35316,164802,40849,46068,20108,31019,21143,86509,30157,109715,96440,15632,34535,57890,184123,70566,10903,91432,40910,80338,48686,16760,43828,30539,156387,49155,44463,29768,10465,11837,118519,47416,40748,100344,56345,10480,57788,33624,29867,79173,29384,13366,144579,26112,122242,21448,51850,79949,18434,167511,52619,45447,51628,37906,110348,367399,128745,137783,42741,25050,324064,59022,36337,56041,125002,183442,17809,168752,44979,15181,25950,79816,62705,116683,43472,21342,46135,34273,42814,103076,28855,18972,47273,92457,28782,33225,648772,14256,52011,3529472,48315,27275,34529,28268,35244,33152,12739,61875,52851,24591,42663,16886,191288,52453,24220,69914,47654,53755,75479,111033,46019,17321,118819,14343,79203,50803,389899,189148,100703,116930,97929,42989,28817,120904,107269,18245,37030,218915,61234,55014,20013,15798,32366,29440,40558,28724,109673,23783,43240,101800,85248,59865,95850,48227,78881,1234826,47831,55666,478845,94258,22423,188559,44524,51898,62755,1236427,34420,20255,1065663,61754,561674,162882,62613,25325,166909,31198,81857,17420,195248,192487,106656,80319,16461,73795,203334,37522,228138,40641,146956,15027,20962,55712,11208,89897,19335,42613,62984,35387,208562,128792,60536,94774,174426,31265,38120,60132,44412,27940,25260,66690,50937,108520,43059,91426,97609,176730,44761,76590,78408,142409,50157,24574,38651};

results = Table[0,{i,221},{j,31}];
answer = Table[0,{i,1},{j,31}];

answer=answer[[1]];
For[j=1,j<=221,j++,
results[[j,1]]=N[CDF[BinomialDistribution[pop[[j]],0.00014883],0]]
For[i=2.5,i<=75,i=i+2.5,Print[i];
    For[j=1,j<=221,j++,
        results[[j,i/2.5+1]]=
N[CDF[BinomialDistribution[pop[[j]],0.00014883],i*pop[[j]]/100000]]
    ]
answer=answer/221
ListPlot[answer,PlotJoined->True]
2.3 Definition of the Observed Distributions

The observed cancer mortality rates distributions are the histograms for the cancer mortality data that are explained in Section 4. These histograms provide the number of towns that exhibit a range of cancer mortality rates for appropriately chosen intervals. The histograms are manipulated in Excel™.

2.4 Comparison of the Chance Expectation and Observed Distributions

2.4.1 Calculation of D*, Application of the K-S Test for Differences Between Distributions

Once the distribution expected by chance is calculated and data to define the observed distribution collected, the distributions were compared to determine whether or not they are significantly different. If significantly different, this would be evidence inherited and/or environmental risk factors have influenced community cancer mortality. A suitable test for significance between two such distributions was the Kolmogorov-Smirnov test [40]. This test (recommended by Dr. Stephan Morgenhaler) is based on the differences between a hypothesized cumulative distribution function and an observed cumulative distribution function. The K-S test was, in fact, chosen for these studies as the sample test statistic is independent of the nature of the distributions [41-43].

To perform the Kolmogorov-Smirnov comparison the cumulative distributions are first calculated for the observed and expected histograms over the range of cancer rates. The differences between the two cumulative distributions are calculated over their entire range. The test statistic, $D^*$, is defined as the maximum value of the differences between the two cumulative histograms. This test statistic, $D^*$, is two-tailed and as defined, Equation 3. Figure 7 provides an illustration of the two cumulative distributions and the hypothetical maximum difference measured as $D^*$. The data used for this example is for
Figure 7: Illustration of the cumulative distributions of the observed and expected by chance and the determination of the D* Statistic for the Kolmogorov-Smirnov (K-S) test.

The data used to create these example distributions is from the colon cancer mortality for European American males (EAM) aged 65-84 for the 251 communities in Massachusetts.
colon cancer mortality for European American males (EAM) aged 65-84 for the 251 communities in Massachusetts which was not significantly different from the chance expectation. Significance is determined by comparing $D^*$ with a critical value $\varepsilon^*$. The calculation for $\varepsilon^*$ is described in the next section.

$$D^* = \max [ | F_1(x) - F_2(x) | ]$$  \hspace{1cm} \text{Equation 3}$$

where,

$F_1$ represents the cumulative values for the observed distribution

$F_2$ represents the cumulative values for the distribution expected by chance

2.4.2 Comparison of $D^*$ and the Critical Value $\varepsilon^*$

This $D^*$ statistic is then compared with a critical value, $\varepsilon^*$, for a desired confidence level of interest, $p$. Two distributions are concluded to be significantly different when the measured $D^*$ value is greater than or equal to the reference value for a desired degree of significance as in Equation 4 below [44].

$$p = \text{Prob} \ (D_n^* \leq \varepsilon^*)$$  \hspace{1cm} \text{Equation 4}$$

where,

$p = \text{the confidence level of interest} \ (p = 1 - 2\alpha \text{ for the two-tailed test})$

(e.g. for a desired confidence level of 95%, $p$ is 0.95 and $\alpha$ is thus 0.025)

This critical value, $\varepsilon^*$, is dependent on the sample size, $n$, the number of communities in the distribution. The equations used to calculate the values for $\varepsilon^*$ are given below. Since the values for $n$ are quite large (>100), tables of the reference values, $\varepsilon^*$, for the desired confidence levels were not available. However, the equations were found to calculate
these values based on \( n \) for a desired confidence level [44]. This is given in the following equation, Equation 5.

\[
\varepsilon^*(n, \alpha) = \bar{\varepsilon}(n, \alpha) - 0.16693n^{-1} - A(\alpha)n^{-3/2} \tag{Equation 5}
\]

where,

\[
\varepsilon^* = \text{value used for comparison to } D^\ast \text{ for } n > 20
\]

\( n \) = number of data values in the distributions

\( \alpha \) = the desired probability of observing a chance occurrence for each comparison

(where \( p = 1 - 2\alpha \) for the two-tailed test)

\[
\bar{\varepsilon}(n, \alpha) = \sqrt{\frac{\ln \frac{1}{\alpha}}{2n}}
\]

\[
A(\alpha) = 0.09037 \left(-\log_{10} \alpha\right)^{3/2} + 0.01515\left(\log_{10} \alpha\right)^2 - 0.08467 \alpha - 0.1143
\]

The equation for \( A(\alpha) \) is an approximation and was found to be increasingly accurate for increasing \( n \) and decreasing \( \alpha \) [44] as is applicable to these analyses. The \( D^\ast \) statistic and the \( \varepsilon^* \) critical value were calculated and evaluated for each pair-wise comparison of the expected and observed distributions.

In this case each cancer analyzed for each cohort represents one pair-wise comparison. As many cancers and cohorts will be analyzed it was necessary to also correct for the number of comparisons. The confidence level suitable for a single pair-wise comparison was not suitable for such a set of comparisons. Using Tippett's formula for continuous distributions [45], given below in Equation 6, \( \gamma \), the probability of observing at least one significant comparison is a function of \( n \), the number of comparisons (or trials), and \( \alpha \), the probability that each trial is significant. Therefore the probability of finding at least one significant outcome for 100 trials when the probability that a single trial is significant is 0.05 is 99.4%. In fact, the number of trials that would be found to be significant by
chance would be 5 or $\alpha*n$ (although the actual number would be slightly less due to the discontinuous nature of the binomial distribution).

$$\gamma = 1 - (1 - \alpha)^n \quad \text{Equation 6}$$

where,

$\gamma$ = the probability of observing at least one significant outcome

$n$ = the number of trials (in this case the number of communities)

$\alpha$ = the probability an individual trial is significant

Therefore, more stringent statistical levels are required for studies with multiple trials to decrease the probability of chance occurrence. The Bonferroni method is used to correct for the problem of multiple comparisons [46-47]. The Bonferroni inequality may be defined as follows:

$$1-(1-\alpha/n)^n \leq \alpha \quad \text{Equation 7}$$

Therefore, the desired overall probability of chance occurrence of $\gamma$, or 0.05, can simply be divided by the number of communities in a set.

$$\alpha \approx \gamma / n \quad \text{Equation 8}$$

This adjustment reduces the probability that an individual comparison is significant by chance from 0.05 to 0.05 divided by the number of trials for a one-tailed test. This Bonferroni approximation is used to adjust the desired confidence level for each pairwise comparison of distributions in the calculation of the critical value, $\epsilon^*$. Therefore, to calculate the critical value $\epsilon^*$ Equation 5 is used, shown again below, with the value of $\alpha$ (the desired probability of observing a chance occurrence for each trial) divided by the number of trials in each set. Therefore, the overall confidence level for the set of trials
remains 95%. Also note that with the use of the two-tailed test 0.025 and not 0.05 is divided by the number of trials to define $\alpha$.

$$
\varepsilon^*(n,\alpha) = \bar{\varepsilon}(n,\alpha) - 0.16693n^{-1} - A(\alpha)n^{-3/2} \quad \text{Equation 5}
$$

where,

$\varepsilon^* = $ value used for comparison to $D^*$ for $n > 20$

$n = $ number of data values in the distributions

$\alpha = $ the desired probability of observing a chance occurrence for each comparison

(where $p = 1 - 2\alpha$ for the two-tailed test)

$$
\bar{\varepsilon}(n,\alpha) = \sqrt{\frac{\ln(\frac{1}{\alpha})}{2n}}
$$

$$
A(\alpha) = 0.09037 \left(-\log_{10}\alpha\right)^{3/2} + 0.01515\left(\log_{10}\alpha\right)^2 - 0.08467\alpha - 0.1143
$$
3. Historical and Geographical Variations

3.1 Historical Variations

Although cancer accounted for about 6% of all deaths at the beginning of the last century it now accounts for over 24% of all deaths in the United States [10-16]. Historical changes in cancer risk are primary evidence that non-genetic factors influence carcinogenesis. These changes may be due to medical improvements, socioeconomic factors as well as environmental factors.

The age-specific mortality rates by birth year cohorts for EAM and EAF for four major forms of cancer (lung cancer, stomach cancer, breast cancer and prostate cancer) in the United States are again shown in Figures 8 to 11 beginning with lung cancer. Figures 8 to 10 were shown previously in the introduction. Lung cancer mortality rates, Figure 8, have been dramatically increasing starting with the birth cohort of the 1820s. Some of the early increase may be due to improved diagnostic accuracy; however, the dramatic increases from the birth cohort of the 1870s for EAM and from the birth cohort of the 1910s for EAF are due to the fraction of the United States population which smoked cigarettes [48]. Although the changing historical picture for lung cancer mortality in the United States can be explained by a known cause, reasons for changes in historical mortality rates are unknown for other common mortal cancers.

Figure 9 shows the mortality rates for stomach cancer on an age- and birth year cohort-specific basis. Unlike lung cancer, which has increased, mortality from stomach cancer has been steadily decreasing in the United States for both EAM and EAF. Over the last approximately 100 years the maximum mortality rates for each birth cohort have decreased more than five-fold. The reason for this is yet unknown, although the changes may be due to medical practices, changes in hygiene and sanitation, or dietary factors. Devesa, et. al., 1987 suggested the decreases may be due to changes in dietary factors such as food preservation [49]. Additionally, risk was observed to be increased in lower socioeconomic groups [50], and rates among migrants are like those of their adopted
country rather than their country of origin [51].

Not all forms of cancer have exhibited dramatic differences as a function of birth year cohort. Figure 10 shows the age-specific mortality rates for breast cancer. The age-specific mortality rates show that the mortality rates have remained fairly constant as a function of history in the U.S. These data contradict the claims of increasing breast cancer risk due to the observed increasing incidence rates from incidence data collected by the Surveillance, Epidemiology, and End Results (SEER) program [52]. The discrepancy between the mortality and incidence data for breast cancer most likely is due to inaccurate diagnoses. Several studies have pointed to overestimates in breast cancer incidence due to diagnostic inaccuracies [53]. Likewise, data for prostate cancer, Figure 11, show increasing mortality rates for the oldest age groups, but are remarkably constant for the ages less than 80 years for the birth cohorts after the 1850s. This may also be due to changes in diagnosis.
Figure 8: United States age-specific mortality rates for lung cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).

$\text{Obs}(h,t)$ per 100,000 is the observed mortality rate per 100,000 for those born in the time period, $h$, who died during the age period, $t$. 
**EAM**

- Birth Year:
  - 1820s
  - 1830s
  - 1840s
  - 1850s
  - 1860s
  - 1870s
  - 1880s
  - 1890s
  - 1900s
  - 1910s
  - 1920s
  - 1930s
  - 1940s
  - 1950s
  - 1960s
  - 1970s
  - 1980s

**EAF**

- Birth Year:
  - 1820s
  - 1830s
  - 1840s
  - 1850s
  - 1860s
  - 1870s
  - 1880s
  - 1890s
  - 1900s
  - 1910s
  - 1920s
  - 1930s
  - 1940s
  - 1950s
  - 1960s
  - 1970s
  - 1980s
Figure 9: United States age-specific mortality rates for stomach cancer as a function of birth cohort for European American males (EAM) and European American females (EAF).

$\text{Obs}(h,t)$ per 100,000 is the observed mortality rate per 100,000 for those born in the time period, $h$, who died during the age period, $t$. 
Figure 10: United States age-specific mortality rates for breast cancer as a function of birth cohort for European American females (EAF).

\( \text{Obs}(h,t) \) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
Figure 11: United States age-specific mortality rates for prostate cancer as a function of birth cohort for European American males (EAM).

Obs(h,t) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
3.2 Geographical Variations

In addition to historical changes in the risk of cancer mortality are changes that have been observed on a geographical basis. These geographical areas have been worldwide, regionally, statewide and between urban and rural areas in the United States. Geographical differences may be due in part to historically changing risk factors that are not yet homogeneously distributed throughout the United States population. Since the introduction of new technologies and products are generally first to urban populations. Risk from these factors will be seen to affect these populations prior to the dissemination through the rest of the population. Therefore observing historical and geographic cancer rate differences may be evidence for the introduction of a risk factors to the population. However, unlike historical changes, geographical differences may also be due to genetic risk factors in addition to medical, socioeconomic and environmental factors. Therefore, variations found in cancer rates on a large-scale may also be subsequently found at the community level.

A study in 1990 of 25 major forms of cancer worldwide showed differences in the forms of cancer prevalent in developed versus developing nations [54]. For instance, in developed countries the prevalent forms of cancer were lung cancer, colon cancer, breast cancer, rectal cancer and prostate cancer. While in developing countries cancer of the cervix and esophageal cancer were most prevalent. Specific comparisons between countries have also shown significant differences between cancer mortality rates. Data for the United States and data compiled by Jose Marquez for Japan beginning in 1950 were used to compare the age-specific mortality rates differences between the two countries for the death years 1950-1991. These age-specific mortality rates curves are shown in Figures 12 and 13 for breast cancer and colon cancer in males, respectively. From Figures 12 and 13 it is obvious that the age-specific mortality rates for Japan are consistently lower than the United States. The reason for these differences are currently unknown and may include genetic, medical, socioeconomic and/or environmental differences. Although all of these risk factors may be relevant to cancer risk between nations, there is evidence that the risk factors cannot only be genetic in nature. Studies of
Figure 12: United States (EA) versus Japan mortality rates for breast cancer as a function of age for the death years from 1950-1991.

Obs(h,t) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
Breast Cancer EAF vs. JF
Birth Decade 1890s

- United States
- Japan
Figure 13: United States (EA) versus Japan mortality rates for colon cancer in males as a function of age for the death years from 1950-1991.

Obs(h,t) per 100,000 is the observed mortality rate per 100,000 for those born in the time period, h, who died during the age period, t.
Colon Cancer EAM vs. JM
Birth Decade 1890s

- United States
- Japan
cancer risk in immigrant populations have shown that the adopted country rather than the country of origin dictated cancer risk [51].

Other geographic variations have been found within the United States itself. There have been many studies of regional cancer incidence and mortality rates that have shown differences throughout the country. In New England cancer mortality rates were found to be consistently higher than national averages for the period of 1950 to 1975. This was true for male cancers, while female cancer mortality rates, shown to be decreasing throughout the United States, decreased less rapidly in New England [55]. County-specific mortality rates from 1950-1969 for bladder cancer showed cancer rates increased with urbanization. The study also showed regional differences with regard to bladder cancer mortality rates with the northeast exhibiting higher rates for both genders in both rural and urban areas. Outside of the northeast bladder cancer mortality rates were found to be consistently higher in urban areas. Among non-whites the regional differences were small [56]. Like bladder cancer breast cancer mortality has also exhibited variations geographically. Mortality data for 1987 to 1991 showed a large difference in mortality rates nationwide, and again mortality rates in the northeast were found to be higher than the rest of the country for breast cancer [57]. Additionally there was also evidence that for urban areas breast cancer was higher as well [57]. The most comprehensive county-specific study performed by Devesa, et. al., 1999 shows the county-specific mortality rates for the years 1950-1994 in the United States [58]. The maps of these data clearly show differences on a regional and statewide basis for many cancers throughout the country.

Although there is evidence that mortality rates are significantly different among regions of the country, the reasons for these differences are unknown. It can be hypothesized that these differences may be due to varying medical care practices, access and utilization. These differences may also reflect changes in risk that occurred through the inhomogeneous introduction of a risk factor to the population. For instance, it is known that smoking is the primary risk factor for lung carcinogenesis [1-2]. Lung cancer mortality rates have been shown to have changed historically concordant with the use of
cigarettes. Therefore an unequal distribution in smoking prevalence rates among the population due to an inhomogeneous introduction of cigarettes would not be unexpected. From the national county-specific lung cancer data it was found that lung cancer mortality was significantly higher in the South, especially for white males [58]. Smoking was found to correlate with the high lung cancer rates as smoking rates were also high in the South in a study by Marcus, et. al., 1989 [59].

Smoking has not only been shown to vary on a regional basis but on a population density basis as well. A study by Haenszel, et. al., 1955 noted higher smoking rates in urban versus rural areas [60]. Likewise, data from the Behavioral Risk Factor Surveillance System (BRFSS) survey obtained through the Massachusetts Department of Public Health (MA DPH) showed that the percentage of current smokers residing in urban communities were slightly higher than for rural communities. These data represented the survey years from 1993 to 1997 and are shown in Figure 14. Although the number of respondents in the data set were too few to determine significance, these data suggest that smoking prevalence was higher in urban than rural communities in Massachusetts. This supports the study by Haenszel, et. al., 1955.

Smoking is but one example of a risk factor that varies throughout the country. There are many other environmental factors that may vary. These risk factors, if any, are as of yet unknown. Therefore to determine whether or not communities exist that are influenced by such risk factors it is expected that any variations in these factors will alter the distributions of cancer rates at the community level. With the evidence of historical and geographic differences in cancer rates in the United States, these differences may also be witnessed at the community level and provide evidence that external factors influence community cancer risk.
Figure 14: Age-specific smoking prevalence comparison between urban and rural communities in Massachusetts.

Data were from the BRFSS Survey from the MA DPH for the years 1993-1997.

The black diamonds represent the percent respondents who smoked in urban communities. The grey triangles are for those in rural communities. The black and grey lines are the smoothed averages through their respective data points.
4. Data

4.1 Acquisition, Compilation and Organization

To perform the community analyses mortality and population data were obtained. Mortality data were obtained as computerized death records. The records were informative for the decedents' age (in single-year age groups), gender, ethnicity, underlying cause of death (according to the International Classification of Disease (ICD) codes), and place of residence at time of death. Due to confidentiality restrictions that varied from state to state, mortality data were obtained for each state on an individual basis after applying for access to the registry data and completing confidentiality agreements. The data obtained and analyzed represented the following death years: 1969-1995 for Massachusetts, 1970-1993 for California, 1970-1997 for Florida and New York, 1979-1998 for Pennsylvania and 1972-1996 for Texas. These data were obtained through the Public Health Departments for each of the aforementioned states with the exception of California. Data for California were obtained from the National Center of Health Statistics (NCHS). These data were not informative for places of residence with a population below 10,000 persons as a population threshold was applied by the NCHS to the data sets after receipt from the California Department of Public Health. Contacts to obtain these data as well as the population data are provided in the Appendix.

Together these populations represent approximately one third of the United States' population. These mortality data were obtained as computerized text files on three types of storage media: data reel tapes, data cartridges and CD-ROM. The text files obtained originally on the storage media of reel tapes and cartridges were downloaded from a mainframe computer on site at MIT to a desktop Macintosh computer using the MITVMA system and Fetch™ (ftp software). All death records are subsequently stored as text files.

The population data were obtained for the six states by age (single-year age groups), gender and ethnicity for all available communities. These data files were obtained from
the United States Census Bureau Division of Special Tabulations as text files for the census years 1970, 1980 and 1990 on a community basis. Additionally, state totals and large community population values were obtained through Census Bureau publications for the same census years and age (5-year totals) and ethnicity/gender cohorts. Population data are also stored on a desktop Macintosh computer. The Census Bureau suppressed population data for a minority of communities due to confidentiality restrictions, and the subsequent data presented are for the subset of the communities in each state with the exception of studies that utilize state averages.

Matlab™ programs were written for the state death record data sets to read and sum the desired data. These programs were unique to each data set, but generally, all programs served to read each record and categorize said record according to the age, gender, ethnicity, ICD code and place of residence for each decedent. Other programs were written to sum the data of interest according to specific cancers and age and ethnicity/gender cohorts. The data were first organized into matrices for each cancer (21 major forms of adult and 6 pediatric) and community by single-age (rows) and ethnicity/gender groups for each death year (columns). An example of these programs, which manipulated the data into matrices is given as Program 2 on the following page. Data organized in these matrices could be subsequently “read” by other Matlab™ programs written to extract data according to the specific categories for study. The data were summed according to the age groups (0-19, 65-84 and ≥ 85 year olds) and ethnicity/gender groups (EAM, EAF, NEAM, and NEAF) for the cancers of interest for each community. An example of these programs is given as Program 3. The adult cancers were lung cancer, colon cancer, breast cancer, prostate cancer, pancreatic cancer, Non-Hodgkin’s lymphoma (NHL), stomach cancer, ovarian cancer, kidney cancer, leukemia, multiple myeloma, esophageal cancer, bladder cancer, uterine and cervical cancer, CNS cancer, liver cancer, rectal cancer, melanoma, gallbladder cancer, laryngeal cancer and pharyngeal cancer. The 6 pediatric cancers were CNS cancer, leukemia, NHL, thyroid and endocrine cancer, connective and soft tissue cancer and bone cancer.
Program 2: Example of a Matlab™ program used for mortality data matrix creation.

... Implies a redundancy in the program that pertains to a listing of communities.
fid=fopen('CA7097.txt.m','r');
F='asd';
load alameda
load albany
load alhambra

load yorbalinda
load yubacity
load state
d=1;
while (not(isempty(F)))
    F=fscanf(fid,'%c',56);
    cause=str2num(F(38:40));
    cause2=str2num(F(41));
    deathyear=str2num(F(3:4));
    county=str2num(F(8:10));
    city=str2num(F(11:13));
    gender=str2num(F(28));
    ethnicity=str2num(F(31));
    age=str2num(F(23:25));
    resstate=str2num(F(6:7));
    if ( ((age<121) | (age>199)) & (deathyear<98) & (deathyear>69) & (age<700) & (cause>171) & (cause<173) & (not(isempty(cause))) & (not(isempty(age))) & (ethnicity>0) & (ethnicity<3) & (not(isempty(ethnicity))) & (gender>0) & (gender<3) & (not(isempty(gender))) & (resstate==5) & (not(isempty(city))) & (not(isempty(county)))
        if age>199
            age=0;
        end
        if ethnicity==1
            if gender==1
                R=1;
            else
                R=2;
            end
        else
            if gender==1
                R=3;
            else
                R=4;
            end
        end
    end
    row=age+1;
column=4*(97-deathyear)+R;
state(row,column)=state(row,column)+1;
if ((city==3) & (county==1)) | ((city==1) & (county==1))
alameda(row,column)=alameda(row,column)+1;
end
if ((city==6)&(county==1))|((city==2)&(county==1))
albany(row,column)=albany(row,column)+1;
end
if ((city==9)&(county==19))|((city==3)&(county==19))
alhambra(row,column)=alhambra(row,column)+1;
end

if ((city==657)&(county==57))|((city==253)&(county==57))
woodland(row,column)=woodland(row,column)+1;
end
if ((city==660)&(county==30))|((city==254)&(county==30))
yorbalinda(row,column)=yorbalinda(row,column)+1;
end
if ((city==663)&(county==51))|((city==255)&(county==51))
yubacity(row,column)=yubacity(row,column)+1;
end

d=d+1
end
fclose(fid);
Program 3: Example of a Matlab™ program used for mortality data extraction from matrices.

- Implies a redundancy in the program that pertains to a listing of communities.
load alameda
load albany
load alhambra
.
.
load yorbalinda
load yubacity
load state
californiadeath = zeros(222,1);
for column = 2:4:108
n=1;
death = alameda(:,column);
californiadeath(n) = californiadeath(n) + sum(death(66:86));
n=n+1;
deaht = albany(:,column);
californiadeath(n) = californiadeath(n) + sum(death(66:86));
n=n+1;
deaht = alhambra(:,column);
californiadeath(n) = californiadeath(n) + sum(death(66:86));
.
.
n=n+1;
deaht = yorbalinda(:,column);
californiadeath(n) = californiadeath(n) + sum(death(66:86));
n=n+1;
deaht = yubacity(:,column);
californiadeath(n) = californiadeath(n) + sum(death(66:86));
n=n+1;
deaht = state(:,column);
californiadeath(n) = californiadeath(n) + sum(death(66:86));
end
deads = zeros(222,1);
for i=1:222
    deaths(i) = californiadeath(i);
end
deads
Data from the computerized Census Bureau data files were similarly read and summed with Matlab™ programs according to the same age, ethnicity/gender and community designations. An example of a population extraction program is given Program 4. These programs created data files of the population values for each community and census year by single-year age and ethnicity/gender groups. These data were then subsequently used to sum the data for the age and ethnicity/gender groups for each community and census year to correspond with the mortality data. An example of these programs is given in Program 5.

The data from each census year were then used to estimate populations for the interim years using linear interpolation between census years and extrapolation for data years subsequent to the 1990 census. These interpolated or extrapolated population values were used in combination with the number of deaths from all causes from the previous period to provide an estimate for the age and ethnicity/gender populations. For instance, the population for those aged 5 for 1971 was estimated using the interpolated value for 6 year olds in 1972 and the total number of deaths for 5 year olds in 1971. This method was used for all communities’ populations. This equation is given below as Equation 9.

\[ \text{pop} (a, y) = \text{intpop} (a+1, y+1) + \text{deaths} (a, y) \quad \text{Equation 9} \]

where,

\[ \text{pop}(a, y) = \text{the desired population estimate for an age group, } a, \text{ in year } y \]
\[ \text{intpop}(a+1, y+1) = \text{the linearly interpolated population estimate for age } a+1 \text{ in year } y+1 \]
\[ \text{deaths}(a, y) = \text{number of deaths for the desired population at age } a \text{ and year } y \]

The deaths, along with these corresponding population values, were then subsequently manipulated with Excel templates arranged specifically for each unique data set to calculate the observed mortality rates, Obs(h,t), according to Equation 10.

\[ \text{Obs}(h,t) = \frac{\# \text{ of persons born in year } h \text{ who died at age } t}{\# \text{ persons born in year } h \text{ who are still alive at age } t} \quad \text{Equation (10)} \]
Program 4: Example of a Matlab™ program used for data extraction for the population files.

. Implies a redundancy in the program that pertains to a listing of communities.
fid=fopen('90CApop2.m','r');
F='asd';
load alameda
load albany
load alhambra
.
.
load yorbalinda
load yubacity
load state
d=1;
age=0;
while (not(isempty(F)))
    if (age==100)
        F=fscanf(fid,'%i',5);
age=F(3);
    else
        F=fscanf(fid,'%i',7);
town1=F(1);
town2=F(2);
age=F(3);
EAM=F(4);
EAF=F(6);
NEAM=F(5);
NEAF=F(7);
row=age+1;
end
if (age<101)
if (((town1==6001)&&(town2==10))
alameda(row,9)=alameda(row,9)+EAM;
alameda(row,10)=alameda(row,10)+EAF;
alameda(row,11)=alameda(row,11)+NEAM;
alameda(row,12)=alameda(row,12)+NEAF;
end
if (((town1==6001)&&(town2==20))
albany(row,9)=albany(row,9)+EAM;
albany(row,10)=albany(row,10)+EAF;
albany(row,11)=albany(row,11)+NEAM;
albany(row,12)=albany(row,12)+NEAF;
end
if (((town1==6037)&&(town2==25))
alhambra(row,9)=alhambra(row,9)+EAM;
alhambra(row,10)=alhambra(row,10)+EAF;
alhambra(row,11)=alhambra(row,11)+NEAM;
alhambra(row,12)=alhambra(row,12)+NEAF;
end
.
.
if ((town1==6113)&(town2==3150))
woodland(row,9)=woodland(row,9)+EAM;
woodland(row,10)=woodland(row,10)+EAF;
woodland(row,11)=woodland(row,11)+NEAM;
woodland(row,12)=woodland(row,12)+NEAF;
end
if ((town1==6059)&(town2==3169))
yorbalinda(row,9)=yorbalinda(row,9)+EAM;
yorbalinda(row,10)=yorbalinda(row,10)+EAF;
yorbalinda(row,11)=yorbalinda(row,11)+NEAM;
yorbalinda(row,12)=yorbalinda(row,12)+NEAF;
end
if ((town1==6101)&(town2==3185))
yubacity(row,9)=yubacity(row,9)+EAM;
yubacity(row,10)=yubacity(row,10)+EAF;
yubacity(row,11)=yubacity(row,11)+NEAM;
yubacity(row,12)=yubacity(row,12)+NEAF;
end
d=d+1
end
fclose(fid);
Program 5: Example of a Matlab™ program used for population data extraction by desired age group.

Imply a redundancy in the program that pertains to a listing of communities.
load alachua
city
load alford
town
load altamonte
springs
city
load yankeetown
town
load zephyrhills
city
load zolfos
springtown

\[ \text{californiapop} = \text{zeros}(221,1); \]
\[ \text{column} = 12 \]
\[ n=1; \]
\[ \text{pop} = \text{alameda}(:, \text{column}); \]
\[ \text{californiapop}(n) = \text{californiapop}(n) + \text{sum}(\text{pop}(66:86)); \]
\[ n=n+1; \]
\[ \text{pop} = \text{albany}(:, \text{column}); \]
\[ \text{californiapop}(n) = \text{californiapop}(n) + \text{sum}(\text{pop}(66:86)); \]
\[ n=n+1; \]
\[ \text{pop} = \text{alhambra}(:, \text{column}); \]
\[ \text{californiapop}(n) = \text{californiapop}(n) + \text{sum}(\text{pop}(66:86)); \]
\[ \]
\[ n=n+1; \]
\[ \text{pop} = \text{yorba} \text{linda}(:, \text{column}); \]
\[ \text{californiapop}(n) = \text{californiapop}(n) + \text{sum}(\text{pop}(66:86)); \]
\[ n=n+1; \]
\[ \text{pop} = \text{yubacity}(:, \text{column}); \]
\[ \text{californiapop}(n) = \text{californiapop}(n) + \text{sum}(\text{pop}(66:86)); \]
\[ n=n+1; \]
\[ \text{pop} = \text{state}(:, \text{column}); \]
\[ \text{californiapop}(n) = \text{californiapop}(n) + \text{sum}(\text{pop}(66:86)); \]
\[ \text{pops} = \text{zeros}(222,1); \]
\[ \text{for } i=1:222 \]
\[ \quad \text{pops}(i) = \text{californiapop}(i); \]
\[ \text{end} \]
\[ \text{pops} \]
4.2 Accuracy

Although some investigators have questioned the value of mortality data due to the hypothesized errors in recording forms of death [61], there are several reasons to use cancer mortality data instead of incidence data. Mortality data are far more accessible than incidence data and are available for any area of the country. Death records have been recorded regularly for over 100 years in the United States. Additionally, these records were computerized beginning about the year 1968 for the majority of states. This eases the collection and manipulation of the data. Incidence data contrarily are available for a small number of registries and only for the last few decades through the SEER (Surveillance, Epidemiology, and End Results) program, which began in 1973 [52 and 62]. There is only one state that has had its own long-standing cancer registry: Connecticut since the 1930s [63]. Additionally, current SEER registries only include a few complete states as some of the registries are confined to urban areas. The current SEER registries are as follows: Alaska, Atlanta, Rural Georgia, Arizona, Connecticut, Detroit, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, San Francisco-Oakland, San Jose-Monterey, Los Angeles, Remainder of California, Seattle-Puget Sound, Utah [62]. These registries were added at varying times since 1973. Kentucky, Louisiana, New Jersey and the Remainder of California were added recently in 2000.

Not only are mortality data more plentiful than incidence data, there have also been several studies that have shown that death record data are fairly accurate. Accuracy of death records was assessed and found to be in good agreement with medical records [64-65]. Cancer deaths in particular have been shown to be recorded with more accuracy than other forms of death [66]. It has also been shown that demographic data from death certificates are also reasonably accurate [67-69]. Likewise, coding for the form of death is instituted on a national level and therefore should be fairly consistent throughout the country [70].
Incidence data, however, has been found to be problematic with regard to diagnostic accuracy. It has been found that breast and prostate cancer incidence data may not be wholly reliable [53 and 71]. Figure 15 provides an example of the problem with data for prostate cancer mortality versus the estimated probability of death from survival data (1-S) over a 70 year period. It is clear that the data are incongruent as survival and prostate cancer mortality data diverge signaling a discrepancy between the incidence and mortality values. Additionally SEER estimates of survival from prostate cancer is now virtually 97% [52]. Even though from Figure 15 we see that mortality has changed little, the decrease in the probability of dying of prostate cancer derived from incidence data shows incidence must be increasing to account for the divergent data. The picture is similar for breast cancer. Breast cancer mortality rates have remained constant historically, Figures 4 and 10. This directly contradicts data showing increasing breast cancer incidence through SEER [52]. Several studies have pointed to overestimates in breast cancer rates through inaccuracies in diagnosis [53]. Therefore, at least for these two cancers is seems mortality data may be more reliable unless incidence can be corrected to account for the inaccurate diagnoses.
Figure 15: A comparison of 1-Survival versus mortality rate values for prostate cancer for EAM, 60 to 64 (62.5) years of age in the United States.
EAM 62.5 year olds

- Mortality rate
- 1-S
- Linear (1-S)

Birth year
5. Analyses of Data

5.1 National versus Six-State Mortality Rates

To illustrate how well the six states represent the nation the death records for the states were summed by age for five major adult and two major pediatric cancers. The data represent the death years 1972-1991 (with the exception of Pennsylvania, which was summed from 1979-1991) and correspond to the data years available for the national data set. The state data were grouped in five-year age intervals for the four ethnicity/gender groups (EAM, EAF, NEAM and NEAF). The mortality rates were then calculated with the corresponding population data available from the United States Census Bureau for the seven forms of cancer and the cohort groups of interest. These age-specific mortality rates were plotted against the same age-specific rates for the nation as an aggregate of the death years. These data are shown in Figure 16 for the following cancers: lung cancer, colon cancer, breast cancer, pancreatic cancer, prostate cancer, leukemia and central nervous system (CNS) cancer.

Using the Log Rank test in conjunction with the Bonferroni inequality the six-state average was found to be significantly different for the seven forms of cancer and the four ethnicity/gender cohorts (p<0.002). The six-state average was higher for the European Americans (EAs) for the older age groups and was lower for the Non-European Americans (NEAs). However, although the age-specific mortality rates are significantly different, the six-state average fairly approximates the national average and remind us of the large-scale differences that exist in cancer mortality rates throughout the nation discussed previously. In general the six states provide a picture of the nation with communities that represent more than approximately one third of the country. These communities also represent the various regions: North, South, East and West and also provide examples of urban, suburban and rural communities.
Figures 16: Composite age-specific mortality rates for the six states and the United States for EAM, EAF, NEAM and NEAM for seven forms of cancer.
Lung Cancer EAM

Lung Cancer EAF

- 6 States
- US

Age (years)
Pancreatic Cancer NEAM

Pancreatic Cancer NEAF

6 States

US
5.2 Distribution Analyses

The community distribution analyses were performed on a state-specific basis for all six states as well as the 21 forms of adult and 6 forms of pediatric cancers for the three age (0-19, 65-84 and ≥ 85 years) and four ethnicity/gender (EAM, EAF, NEAM, NEAF) cohorts. Distributions were analyzed for each cancer and cohort for which sufficient data were available to construct histograms. In a few cases, primarily for the NEA population for the most rare cancers where less than 25 communities had a non-zero result, the distributions could not be constructed and analyzed. The data for each state and cancer were first extracted and summed on a cohort-specific basis for each community. The corresponding populations were also extracted and calculated, and the resulting cohort- and community-specific mortality rates were recorded. These mortality rates were further organized as histograms with intervals of judicious size to appropriately construct distributions appropriate for analysis, see 5.2.3. The expected distributions were subsequently calculated as described previously to correspond to each of the observed histograms. These expected distributions account for the variations in the mortality rates expected due to chance. These chance expectations are calculated with the average mortality rate for the state, the populations for the cities/towns and the binomial probabilities of observing a range of mortality rates for each set of cities/towns.

5.2.1 Consideration of Weighted and Unweighted Mortality Rate Averages

In most cases the average used to calculate the statistically expected distribution was the state average mortality rate. However, in some instances the state average was not representative of the average for the observed set of communities. This was not wholly unexpected as the community data available for analyses were a subset for each state. The available communities were more likely to be more urban in nature as the rules for suppression of confidential data, which are based on the racial distribution of each town, by the Census Bureau were more likely to exclude smaller communities. Additionally, the mortality data available from the NCHS for California were for communities with a
population greater than 10,000 persons. This could create a condition where the state average would be slightly lower than the average for the set of communities analyzed. Conversely, the state average mortality rate and the average for the community distribution may also differ if a large city exhibits a significantly different mortality rate than the rest of the state. An example of this is New York City. An increased or decreased mortality rate for such a sizable city would naturally skew the state average rate and would not accurately represent the actual distribution for the set of communities. In these instances utilizing the state average mortality rate would cause a shift in the expected distributions from the observed instead of the predicted alteration in the distribution shape. Figure 17 below shows an example where the observed and expected distributions exhibited this shifting. The figure shows the observed and expected distributions for stomach cancer mortality rates for the communities in New York for EAM aged 65-84 years. The expected distribution was calculated with the state average mortality rate. In this case, it was found that the state average was different due to the influence of the large urban cities specifically New York City.

For the instances where the use of the state average mortality rate created a shift in the expected distribution from the observed, and consequently an artificial determination of a significant difference, an average rate representative of the observed distribution was used to recalculate the distribution accounting for chance expectation. This average was calculated and tested in two ways. The average was first calculated as the unweighted average of the community mortality rates. This was necessary for the influence of very large cities. Alternatively another average was also calculated and tested. The second alternate average was simply the weighted average for only the communities in the data set rather than the weighted average for all of the communities in the state (the state average). The alternate averages were tested in addition to the state average in cases where the observed and chance distributions were found to be significantly different with the use of the state average. In the example of stomach cancer for EAM aged 65-84, the alternate average (unweighted community average) caused the distributions to no longer be significantly different. This is shown below in Figure 18.
The following is a list of cancer and cohorts for which alternate average cancer mortality rates were used and subsequently were found not to be significantly different with the exception of prostate cancer in New York for those aged ≥ 85 year olds. In this case the distributions exhibited the shifting indicative of a possible urban/rural difference. However, none of the state averages accounted for the shift. This particular case may have been a result of as of yet unknown computational or data error and as a result was not included in the list of distributions found to be significantly different as a result of increased variance in the distributions. This list is given in the following pages. These results indicate an urban/rural difference detected with the Kolmogorov-Smirnov.

CA  Lung cancer NEAF 65-84 year olds  (p<0.00032)
    Colon cancer NEAM, NEAF 65-84 year olds
    Breast cancer for NEAF 65-84 year olds
    Pancreatic cancer NEAM, NEAF 65-84 year olds
    Ovarian cancer for NEAF 65-84 year olds
    Multiple Myeloma NEAM, NEAF 65-84 year olds
    Esophageal cancer NEAM 65-84 year olds
    Bladder cancer NEAM, NEAF 65-84 year olds
    Uterine and Cervical cancer NEAF 65-84 year olds

FL  Lung cancer for EAM ≥ 85 year olds  (p<0.00038)
    Colon cancer for EAM ≥ 85 year olds
    Prostate cancer for EAM 65-84 and ≥ 85 year olds
    Esophageal cancer for EAM 65-84 year olds
    Bladder cancer for EAM ≥ 85 year olds
    Laryngeal cancer for EAM 65-84 year olds

MA  Lung cancer for EAM for 65-84 year olds (p<0.0005)

PA  Lung cancer for EAM ≥ 85 year olds  (p<0.0004)
    Colon cancer for EAM ≥ 85 year olds
    Prostate cancer for EAM ≥ 85 year olds
    Breast cancer for EAF 65-84 and ≥ 85 year olds

NY  Lung cancer for EAM 65-84 year olds  (p<0.00038)
Colon cancer for EAM, EAF 65-84 and EAF ≥ 85 year olds
Prostate cancer for EAM ≥ 85 year olds
Laryngeal cancer for EAM 65-84 year olds
Liver cancer for EAM, EAF 65-84 year olds
Stomach cancer for EAM, EAF 65-84 year olds
Pancreatic cancer for EAM, EAF 65-84 year olds
Breast cancer for EAF 65-84 and ≥ 85 year olds
Leukemia for EAM 0-19 year olds

TX
Lung cancer for NEAF 65-84 year olds  (p<0.00033)
Colon cancer for NEAM, NEAF 65-84 year olds
Prostate cancer for EAM ≥ 85 year olds
Stomach cancer for NEAM 65-84 year olds
Pancreatic cancer for NEAM, NEAF 65-84 year olds
Breast cancer for NEAF 65-84 year olds
Leukemia for NEAM 68-84 year olds
Multiple Myeloma for NEAF 65-84 year olds
Uterine and Cervical for NEAF 68-84 year olds
Figure 17: Observed and expected by chance stomach cancer mortality rate distributions for the 477 New York cities/towns, EAM, 65-84 years of age.
New York Stomach Cancer
EAM, 65-84

Number of Communities

Mortality Rate per 100,000

- Observed
- Numerical Expectation
Figure 18: Observed and revised expected by chance stomach cancer mortality rate distributions for the 477 New York cities/towns, EAM, 65-84 years of age.
New York Stomach Cancer
EAM, 65-84

Number of Communities

Mortality Rate per 100,000

- Observed
- Numerical Expectation
5.2.2 Distribution Analyses Results

Once the expected distributions due to chance variation were calculated, the observed and expected were compared with the K-S method described previously for the 21 forms of adult and 6 forms of pediatric cancer for the three age- and four ethnicity/gender- cohorts. For each state-specific set, the probability level utilized to calculate the values for $\epsilon^*$ were first adjusted by dividing 0.025 with the number of comparisons in each state set. The number of comparisons for each state were approximately the product of the number of cancers, the number of age groups and the number of ethnicity/gender groups used in the comparisons corrected for the number of distributions not analyzed due to fewer than 25 non-zero observations. The comparisons that were determined to be significantly different were recorded. The following is a list of the distributions that were determined to be significantly different with the K-S test for each state.

FL  Lung cancer for EAM, EAF and NEAM for 65-84 year olds  (p<0.00038)
PA  Lung cancer for EAM and EAF 65-84 year olds  (p<0.0004)
TX  Lung cancer for EAM, EAF, NEAM and NEAF 65-84 year olds  (p<0.00033)
     Prostate cancer for NEAM 65-84 year olds
CA  Lung cancer for EAM, EAF, NEAM 65-84 year olds  (p<0.00032)
     Prostate cancer for EAM and NEAM 65-84 year olds
     Breast cancer for EAF 65-84 year olds

For the majority of cancers and cohorts the observed distributions of community mortality rates were not significantly different from what is expected by chance variation alone. The following is again a list of the adult cancers studied: lung cancer, colon cancer, breast cancer, prostate cancer, pancreatic cancer, NHL, stomach cancer, ovarian cancer, kidney cancer, leukemia, multiple myeloma, esophageal cancer, bladder cancer, uterine and cervical cancer, CNS cancer, liver cancer, rectal cancer, melanoma, gallbladder cancer, laryngeal cancer and pharyngeal cancer. The 6 pediatric cancers are as follows: CNS cancer, leukemia, NHL, thyroid and endocrine cancer, connective and soft tissue cancer and bone cancer.
Even though there were a few cancer/cohorts determined to be distributed significantly different from chance, it was striking how closely the distributions expected by chance aligned with the observed distributions. For illustrative purposes, Figure 19 shows the observed and expected distributions for all 21 forms of adult (65-84 year olds) and 6 forms of pediatric (0-19 year olds) cancer for EAM (with the exception of the female specific cancers) for the state of Pennsylvania. Pennsylvania was typical of all six states and also had the largest number of communities analyzed (771). For Pennsylvania EAM, aged 65-84 years lung cancer was statistically significant, and the figure exhibits the predicted result with a broadening of the observed distribution and a lower maximum than the expected. Table 1 lists the D* values for all approximately 800 distribution comparisons and is given immediately after the figures illustrating the observed and expected distributions for Pennsylvania.
Figure 19: Observed and expected by chance distributions for the 21 adult and 6 pediatric forms of cancer for EAM (with the exception of the female specific cancers) for the state of Pennsylvania.
Stomach Cancer EAM
PA

Ovarian Cancer EAF
PA
Table 1: D* test statistic values for all comparisons performed for each state.
<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>California</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer</td>
<td>0.03</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>0.14</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>0.17</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>0.09</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>0.07</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>0.07</td>
</tr>
<tr>
<td>Kidney Cancer</td>
<td>0.08</td>
</tr>
<tr>
<td>NHL</td>
<td>0.09</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.04</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>0.07</td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>0.09</td>
</tr>
<tr>
<td>Uterine and Cervical Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>CNS Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td>0.08</td>
</tr>
<tr>
<td>Rectal Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.04</td>
</tr>
<tr>
<td>Gallbladder Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Laryngeal Cancer</td>
<td>0.07</td>
</tr>
<tr>
<td>Thyroid and Endocrine Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Connective and Soft Tissue Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Bone Cancer</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age Group</th>
<th>California</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-19</td>
<td>0.03</td>
</tr>
<tr>
<td>20-29</td>
<td>0.02</td>
</tr>
<tr>
<td>30-39</td>
<td>0.14</td>
</tr>
<tr>
<td>40-49</td>
<td>0.17</td>
</tr>
<tr>
<td>50-59</td>
<td>0.09</td>
</tr>
<tr>
<td>60-69</td>
<td>0.07</td>
</tr>
<tr>
<td>70-79</td>
<td>0.08</td>
</tr>
<tr>
<td>80-89</td>
<td>0.04</td>
</tr>
<tr>
<td>90+</td>
<td>0.03</td>
</tr>
<tr>
<td>Cancer Type</td>
<td>0-19</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>0.31</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>0.15</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>0.14</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>0.07</td>
</tr>
<tr>
<td>NHL</td>
<td>0.03</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>0.14</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td></td>
</tr>
<tr>
<td>Kidney Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Leukemia</td>
<td></td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>0.08</td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>0.1</td>
</tr>
<tr>
<td>Uterine and Cervical Cancer</td>
<td></td>
</tr>
<tr>
<td>CNS Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Rectal Cancer</td>
<td>0.07</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.06</td>
</tr>
<tr>
<td>Gallbladder Cancer</td>
<td>0.03</td>
</tr>
<tr>
<td>Laryngeal Cancer</td>
<td>0.09</td>
</tr>
<tr>
<td>Pharyngeal Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Thyroid and Endocrine Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Connective and Soft Tissue Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Bone Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>0-19 EAM EAF NEAM NEAF</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>0.09</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>0.11</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>NHL</td>
<td>0.01</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>Kidney Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.09</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>0.06</td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Uterine and Cervical Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>CNS Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Rectal Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>Gallbladder Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.01</td>
</tr>
<tr>
<td>Pharyngeal Cancer</td>
<td>0.01</td>
</tr>
<tr>
<td>Thyroid and Endocrine Cancer</td>
<td>0.01</td>
</tr>
<tr>
<td>Connective and Soft Tissue Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Bone Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Cancer Type</td>
<td>0-19 EAM</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
</tr>
<tr>
<td>Colon Cancer</td>
<td></td>
</tr>
<tr>
<td>Breast Cancer</td>
<td></td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td></td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td></td>
</tr>
<tr>
<td>NHL</td>
<td>0.05</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td></td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td></td>
</tr>
<tr>
<td>Kidney Cancer</td>
<td></td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.02</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td></td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td></td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td></td>
</tr>
<tr>
<td>Uterine and Cervical Cancer</td>
<td></td>
</tr>
<tr>
<td>CNS Cancer</td>
<td>0.08</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td></td>
</tr>
<tr>
<td>Rectal Cancer</td>
<td></td>
</tr>
<tr>
<td>Melanoma</td>
<td></td>
</tr>
<tr>
<td>Gallbladder Cancer</td>
<td></td>
</tr>
<tr>
<td>Laryngeal Cancer</td>
<td></td>
</tr>
<tr>
<td>Pharyngeal Cancer</td>
<td></td>
</tr>
<tr>
<td>Thyroid and Endocrine Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Connective and Soft Tissue Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>Bone Cancer</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Pennsylvania 65+</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>0.092</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>0.01</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Kidney Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.006</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>0.02</td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Uterine and Cervical Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>CNS Cancer</td>
<td>0.005</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td>0.008</td>
</tr>
<tr>
<td>Rectal Cancer</td>
<td>0.008</td>
</tr>
<tr>
<td>Gallbladder Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Laryngeal Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Thyroid and Endocrine Cancer</td>
<td>0.006</td>
</tr>
<tr>
<td>Connective and Soft Tissue Cancer</td>
<td>0.003</td>
</tr>
<tr>
<td>Bone Cancer</td>
<td>0.003</td>
</tr>
<tr>
<td>Cancer Type</td>
<td>0-19</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>0.24</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>0.11</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>0.12</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>0.09</td>
</tr>
<tr>
<td>NHL</td>
<td>0.02</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Ovarian Cancer</td>
<td>0.09</td>
</tr>
<tr>
<td>Kidney Cancer</td>
<td>0.1</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.03</td>
</tr>
<tr>
<td>Multiple Myeloma</td>
<td>0.04</td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>0.06</td>
</tr>
<tr>
<td>Uterine and Cervical Cancer</td>
<td></td>
</tr>
<tr>
<td>CNS Cancer</td>
<td>0.03</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td></td>
</tr>
<tr>
<td>Rectal Cancer</td>
<td>0.02</td>
</tr>
<tr>
<td>Melanoma</td>
<td>0.03</td>
</tr>
<tr>
<td>Gallbladder Cancer</td>
<td>0.05</td>
</tr>
<tr>
<td>Laryngeal Cancer</td>
<td></td>
</tr>
<tr>
<td>Pharyngeal Cancer</td>
<td>0.04</td>
</tr>
<tr>
<td>Thyroid and Endocrine Cancer</td>
<td>0.01</td>
</tr>
<tr>
<td>Connective and Soft Tissue Cancer</td>
<td>0.01</td>
</tr>
<tr>
<td>Bone Cancer</td>
<td>0.02</td>
</tr>
</tbody>
</table>
5.2.3 Histogram Interval Size

To arrange the observed cancer mortality rates for the communities in each state for each of the distribution comparisons it was necessary to choose an interval size for each histogram that allowed for the analysis of a distribution representative of the data. Subsequently the expected by chance probabilities for each set were calculated for the corresponding intervals from this observed histogram. Because of this dependence on the selection of histogram interval size and the resulting number of intervals, the test statistic (and consequently the determination of significance for the comparison of these histograms) may also be dependent on the choice of intervals. Therefore, an experiment was performed to investigate the stability in the measurement of $D^*$ as the number of intervals for the histograms were varied. The data used in this analysis were the mortality rates for all available death years, 1969-1995, for Massachusetts for the following cancers and cohorts: lung cancer for EAM aged 65-84, leukemia for EAM aged 0-19 and breast cancer for EAF aged 65-84.

The histograms for the observed mortality rates were calculated for a varying number of intervals. The expected distributions were subsequently calculated to the same intervals of the observed histograms and the resulting values for the $D^*$ test statistic were recorded. It was found that the test statistic was remarkably constant even with a varying number of intervals. Figure 20, shows the values of $D^*$ that were measured for the comparisons over the range of a number of interval choices. It is shown in this graph that the $D^*$ value varied little when the number of intervals used were between 12 to 22. All of the analyses were performed within this range of interval size.
Figure 20: Values for the D* statistic for the Kolmogorov-Smirnov test as the number of histogram intervals were varied for three cancers and cohorts.
All 800 distribution analyses utilized a number of intervals within the range of 12 to 22 intervals.
5.2.4 Accuracy and Detection of Perturbations of the K-S Method

Although the K-S test is a common method used to compare two distributions there were no prior studies available on its accuracy or its ability to detect variations within data sets. Therefore, it was necessary to perform two studies of the K-S test: a simulation to test its accuracy and an analysis of perturbations detected by the test with a sample data set. First a simulation was performed to determine its accuracy. 1000 randomly generated data sets were compared with the statistical expectation to compare the number of significant occurrences the K-S test would detect with the number expected by chance (5%).

The hypothetical case of 100 towns each with 1000 persons and an average probability of death of 0.05 was used to define the binomial distribution from Equation 1. A Mathematica™ program randomly assigned a number of deaths for each of the 100 towns according to this defined binomial distribution 1000 times. The simulation program is given as Program 6. Each of the 1000 simulations provided the “observed” distributions which were subsequently compared to the distribution defined by the binomial distribution of the starting values. With the standard 95% confidence limits, 5% of these simulations were expected to be significantly different by chance. Thus 50 of the 1000 simulations were expected to be significant with the K-S test. The two-tailed K-S test resulted in 20 distributions to be significantly different than expected. Using the one-tailed K-S test, 42 were found to be significantly different. The two-tailed K-S test was less sensitive than expected; however, the difference in the corresponding test statistic critical values, $\epsilon^*$, were not large enough to significantly change the conclusions drawn from the distribution analyses.
Program 6: Mathematica™ program for the random simulation of the number of deaths assuming a specified binomial distribution.
<<Statistics'DiscreteDistributions`

pop = {1000};
results = Table[0,{i,1},{j,41}];
answer = Table[0,{i,1},{j,41}];
results[[1,1]]=N[CDF[BinomialDistribution[pop[[1]],0.05],0]];
For[i=2,i<=80,i=i+2,Print[i];
results[[1,i/2+1]]=N[CDF[BinomialDistribution[pop[[1]],0.05],i]]]
results
A data set was chosen to illustrate the magnitude of perturbations the K-S test detects in the community distribution analyses. This was determined by perturbing a sample data set and determining which of the perturbations would be detected with the K-S test. Pancreatic cancer for EAM aged 65-84 years for the 251 available communities in Massachusetts was utilized as the test data set. The original data were manipulated for three scenarios. First the data were changed to artificially increase the mortality rates for a set number of small communities. All of these communities had a cohort population of less than 20,000 persons. There were 159 communities that fell into this category. The communities' mortality rates were first increased 50% then 2-fold and finally 5-fold. This was performed for a varying number of these small towns (2, 5, 10, 20 and 40 (for the 2-fold and 5-fold cases only)) chosen at random. These manually changed distributions were then compared to the expected and the maximum differences between the cumulative distributions, D*, were recorded. The data were also changed to artificially increase the mortality rates for the five largest communities. The cohort populations for the five largest communities were greater than 140,000 persons. The communities' mortality rates were first increased 50% then 2-fold and finally 5-fold for the following number of largest cities (1, 3, and 5). In addition to increasing the mortality rates for a number of small towns and cities, the data were also manipulated to test the case where a mix of communities' mortality rates were increased and decreased. The community data were artificially perturbed by increasing and decreasing the observed mortality rates for a randomly chosen number of small-, medium- and large-sized communities. This was done for 10, 20 and 40 communities where the mortality rates were increased and decreased 50%, increased 2-fold and decreased 50%, and increased 5-fold and decreased 75%.

The results for these perturbations are shown in Figure 21. The results from these tests show that the K-S test has the power to detect differences that are due to increases or decreases in the largest communities as well as a mix of increases and decreases among a number of small and large communities. The test detected significant differences for distributions where approximately 40 small communities (<20,000 persons) had increased cancer rates of about five-fold. Additionally, the test was also able to detect
differences resulting from increases for the largest five communities of 50% and a two-fold difference for the sets of largest communities. Likewise differences were easily detected for a random mix of communities with increased and decreased cancer rates. This is the scenario that most closely represents what would be expected if cancer risk factors were to vary throughout communities. A set of about 40 communities were found to be significantly different with increases of two- and five-fold and corresponding decreases of 50% and 75%, respectively. The K-S test, however, would not detect small differences among small communities unless the number of communities that are affected are a great percentage of the overall number of cities/towns.

In general as the expected number of deaths increase in a cohort population with a more common form of cancer or with a larger cohort population size, the sensitivity of detection for the K-S test also is expected to increase. Conversely, as the number of expected deaths from a more rare form of cancer or due to a smaller cohort population size, the sensitivity of the K-S test would decrease. With the results from the distribution analyses, it has been shown that the K-S test is powerful enough to detect variations in lung cancer risk at the community level indicating that the test can detect variations in smoking prevalence within communities. Additionally, the K-S test was also sensitive enough to detect urban/rural differences within the community data sets for a number of cancers for each state.
Figure 21: Perturbations of the Massachusetts pancreatic data set for EAM, 65-84 year olds and the level of detection ($D^*$) with the K-S method.

Case 1: A Number ($n=2, 5, 10$ and $20$) of Small Communities’ Mortality Rates were Increased 50%, 2-fold and 5-fold.

Case 2: A Number ($n=1, 3, and 5$) of the Largest Communities’ Mortality Rates were Increased 50%, 2-fold and 5-fold.

Case 3: A Number ($n=10, 20$ and $40$) of Small to Large Communities’ Mortality Rates are Increased and Decreased 50%, Increased 2-fold and Decreased 50%, and Increased 5-fold and Decreased 75%.
Small Communities (<20,000 persons) with Higher Rates

Largest Communities (>140,000 persons) with Higher Rates
A Mix of Communities with Higher and Lower Rates

D* Statistic

Number of Communities

(\varepsilon^*)

value for significance

△ +/- 50 %
× +2 fold/-50% 
• +5 fold/-75%

140
5.2.5 Analyses for Individual Communities

Although the distribution analyses demonstrate the majority of cancers for communities in six states are distributed by chance, this does not preclude the possibility that there are individual communities within the overall distributions that are significant. From the analyses of the perturbations detectable with the K-S test, a small number of communities which may be significantly different will not be detected in the analyses of the distributions of the cancer mortality rates for the set of communities. In order to examine the significance for individual communities, it is necessary to calculate the binomial probabilities for each community. Once the binomial probabilities are calculated for each community, cancer and cohort, with the observed state average rate, (which coincided with the averages used for the distribution analyses) these probabilities were then compared to a stringent probability limit defined according to the Bonferroni inequality. Therefore, it was necessary to employ a confidence limit that would avoid the possibility that the communities will be determined to be significantly different due to chance as the data set represents a large number of communities for many cancers and cohorts.

The binomial probability for each community was calculated using the observed number of deaths, \( d_k \), in that community for the particular cohort and cancer, the population for the cohort within the community, \( n_k \), and an average rate for the cancer and cohort, \( \rho \).

From Figure 6 shown previously we observed for Massachusetts breast cancer in EAF aged 65-84 years, the ratio of the observed and expected number of deaths for the communities exhibited greater dispersion as the population size decreased. Figure 6 is shown again on the following page. Therefore the probability of observing a several-fold difference in the number of cancers deaths relative to the expected number is grows as the populations size decreases. The equation to calculate these probabilities was given in Equation 1 previously. These probabilities then were compared with the Bonferroni limit defined as 0.025 divided by the number of communities in the data set investigated. This was performed for the communities for all six states for five major adult (aged 65-84) and two pediatric (aged 0-19) cancers for EAM and EAF. The cancers for these analyses are
Figure 6: Ratio of the observed and expected number of cancers deaths for breast cancer in Massachusetts’ 251 cities and towns versus the log of the cohort population size for European American females aged 65 to 84 years in the period 1969-1995.

As the log of the cohort population size decreases the ratio of the observed number of cancers and the expected increases in magnitude.
as investigated previously for the six-state versus national comparisons: lung cancer, colon cancer, breast cancer, prostate cancer, pancreatic cancer, leukemia and CNS cancer. In each set there were 2216 communities representing the six states for each age and ethnicity/gender cohort. Therefore, using the 95% confidence interval in conjunction with the Bonferroni inequality results in a probability level of $1.13 \times 10^{-5}$. Therefore, for the seven cancers and the two ethnicity/gender groups investigated (with the exception of prostate and breast cancer) using this probability limit there is a 1 in 20 chance for each of the 12 sets of data that a significant result will be observed.

The use of the Bonferroni is only an approximation due to the discontinuous nature of the data sets. Therefore, a simulation was performed to determine how the Bonferroni limit would correspond to what would be found using the data set and Monte Carlo simulations. Using the 2216 communities’ population values and an average mortality rate, a Matlab\textsuperscript{TM} program was written to randomly produce a number of deaths for each of the 2216 communities 1000 times according to a defined binomial distribution (Equation 1). An example of this program is given as Program 7. The individual binomial probabilities were calculated for this random number of deaths, and the lowest 50 binomial probabilities were recorded. This corresponds to the 95% confidence limits for the set of all 2216 communities. The largest of which was compared to the Bonferroni approximation. This simulation was performed for two examples by using the average mortality rates for lung cancer for EAM aged 65-84 years and leukemia EAM aged 0-19 years for the communities for each state. The probability level determined from the 1000 simulations for lung cancer was $6.96 \times 10^{-6}$ while for leukemia was $6.88 \times 10^{-5}$. The Bonferroni limit again was $1.13 \times 10^{-5}$. The fold difference between the simulated probability level and the Bonferroni approximation for these cases were between two- and six-fold difference. This would have little effect on the results or conclusions.

For each cancer and cohort investigated it was now possible to identify all communities that fell below this probability level in the set. Remember the choice of this probability
level reduces the number of communities that will be significantly different by chance but does not eliminate chance occurrence altogether. It is also important to note that communities that are significantly different may have a higher or lower than expected number of deaths. The following, Table 2, lists the forms of cancer investigated and the number of instances where a community’s mortality rate fell below this probability level for EAM and EAF. The table also lists the number of significant occurrences that were a result of an increase in the community’s mortality rate. Significant results were circled when consistent with the results from the K-S distribution analyses.
Program 7: Simulation program to test the difference between the simulated probability level and the Bonferroni limit.
load pop.m
load mortrate.m
sims=1000;
cities=2216;
results=zeros(1,sims);
answer=ones(1,sims*0.05);
for i = 680:cities
  i
temppop=pop(i);
temprate=mortrate(i)/100000;
randnum=random('bino',temppop,temprate,1,sims);
results=binopdf(randnum,temppop,temprate);
temp=sort([answer results]);
answer=temp(1:0.05*sims);
end
save answer.m answer -ascii
Table 2: Number of communities for which the observed mortality rates had a probability of less than that defined by the 95% confidence level in conjunction with the Bonferroni limit ($p<1.13 \times 10^{-5}$) for the listed cancers.

For each state, cancer and cohort the number of significant communities are listed as well as the number of those, which are due to increases in the mortality rate. This is indicated as two numbers in the parentheses. The first is the number of total significant communities.

The numbers circled are those which were also identified by the Kolmogorov-Smirnov method of comparing the observed and expected by chance distributions. The numbers enclosed by a box are those which were identified initially with the Kolmogorov-Smirnov test and were subsequently found to be not significantly different with the alternate averages.
<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>California EAM</th>
<th>California EAF</th>
<th>New York EAM</th>
<th>New York EAF</th>
<th>Texas EAM</th>
<th>Texas EAF</th>
<th>Pennsylvania EAM</th>
<th>Pennsylvania EAF</th>
<th>Florida EAM</th>
<th>Florida EAF</th>
<th>Massachusetts EAM</th>
<th>Massachusetts EAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Cancer</td>
<td>67, 43</td>
<td>37, 23</td>
<td>29, 13</td>
<td>8, 6</td>
<td>29, 7</td>
<td>19, 11</td>
<td>19, 14</td>
<td>6, 5</td>
<td>39, 12</td>
<td>16, 8</td>
<td>9, 5</td>
<td>3, 0</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>21, 18</td>
<td>17, 14</td>
<td>8, 8</td>
<td>5, 4</td>
<td>4, 3</td>
<td>7, 4</td>
<td>1, 1</td>
<td>2, 2</td>
<td>11, 10</td>
<td>10, 9</td>
<td>1, 1</td>
<td>1, 1</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>25, 23</td>
<td></td>
<td>0, 0</td>
<td></td>
<td>6, 4</td>
<td></td>
<td>1, 1</td>
<td></td>
<td>14, 4</td>
<td></td>
<td>0, 0</td>
<td></td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>20, 14</td>
<td>(11, 11)</td>
<td></td>
<td></td>
<td>10, 6</td>
<td>(1, 1)</td>
<td></td>
<td>(13, 11)</td>
<td>(0, 0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>7, 6</td>
<td>6, 5</td>
<td>2, 2</td>
<td>1, 1</td>
<td>3, 2</td>
<td>4, 3</td>
<td>0, 0</td>
<td>0, 0</td>
<td>4, 4</td>
<td>4, 4</td>
<td>0, 0</td>
<td>0, 0</td>
</tr>
<tr>
<td>Pediatric Leukemia</td>
<td>0, 0</td>
<td>0, 0</td>
<td>1, 1</td>
<td>0, 0</td>
<td>1, 1</td>
<td>1, 1</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
</tr>
<tr>
<td>Pediatric CNS Cancer</td>
<td>0, 0</td>
<td>0, 0</td>
<td>1, 0</td>
<td>0, 0</td>
<td>1, 1</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
<td>0, 0</td>
</tr>
</tbody>
</table>
For three major forms of cancer the ratios of the observed and expected numbers of cancer deaths were again plotted against the log of the population values for all 2216 communities. This was done for lung, colon and breast cancer for EAM aged 65-84 years with the exception of breast cancer where EAF data were considered. The figures are shown on the following pages as Figure 22. It is again clear that the dispersion increases with decreasing population size. It is also clear that the data represented in this way clearly shows how most of the data points lie within a cone shaped distribution of probable values while it is also easy to directly observe the outliers.
Figure 22: Ratio of the observed and expected number of cancers deaths for lung, colon and breast cancer for all 2216 cities and towns versus the log of the cohort population size for European American males aged 65 to 84 years for all available death years (European American females were considered for breast cancer).

As the log of the cohort population size decreases the ratio of the observed number of cancers and the expected increases in magnitude. It is clear that data presented in this manner clearly shows the expected shape of the distribution of values, which may best be described as a cone. It is also clear from direct observation that outliers exist in each of the graphs with the largest number for lung cancer.
Lung Cancer

Outliers
Colon Cancer

Outliers

log[Cohort Population Size]

Observed/Expected
Breast Cancer

log[CoHORT Population Size]

Observed/Expected

Outliers
6. Discussion

6.1 Distribution Analyses

For the majority of 21 adult and 6 pediatric cancers for the four ethnicity/gender and three age groups for six states in the U.S., the community cancer mortality rates were found to be consistent with what is expected by chance variation alone. This was true after recalculation with an alternate average necessary for several cancers and cohorts due to a difference in the state average and the average for the observed set of communities. These differences indicated an urban/rural effect for many of the cancers and cohorts. Evidence for the urban effect for lung cancer [73-74] as well as for other major forms of cancer has been previously documented [56-57]. After recalculation of these distributions, it was found that the majority did not exhibit significant differences from chance indicating risk factors were limited to the urban/rural effects and did not vary significantly among the communities. There were, however, exceptions. Out of the approximately 800 distributions comparisons, 16 occurred that were significantly different from the expected by chance. Of these 16, 12 occurred for lung cancer.

Lung cancer was the only cancer that consistently diverged from the theoretically expected. The divergence of lung cancer is most likely due to smoking habit practices, which may differ from community to community. It was shown previously that smoking prevalence rates differ significantly on a regional basis [59] and between urban and rural areas [60]. Therefore significant differences were not unexpected for lung cancer. Although there are other factors that may be considered to influence lung cancer risk, smoking is the greatest determinant of risk, and influence of other factors, if any, are most likely small. Quantitative modeling of lung cancer age-specific data has shown that virtually all smokers are at risk for lung cancer eliminating the possibility that familial clustering within communities would influence lung cancer rate variations [48]. These results infer that all individuals are at risk for lung cancer given they smoke. The other factors that may influence the distributions for lung cancer mortality rates at the community level are diagnostic accuracy and the risk of competing causes of death.
Survival rate variations can most likely be eliminated as a cause for the variation of lung cancer at the community level due to the fact that survival from lung cancer is very low, virtually zero [52].

Differences for California were quite anomalous from the observations of the five other states. This may be a result of variations in census counts. The over- and underestimates of these populations due to the racial designation of white or non-white Hispanics may occur in states with a sizable Hispanic population. The change in racial designation occurred with the 1980 census and caused the further delineation of Hispanics in both White and Non-White categories [72 and personal communication with the United States Census Bureau]. This variant inclusion of Hispanics in the White or Non-White racial designations for the Census Bureau population values could subsequently lead to under- or overestimates in the mortality rates as the racial designations for the mortality files did not have this delineation. These under- and overestimates would create variations in both EA and NEA mortality rates with the variations for NEA being more significant as Hispanics are a larger percentage of the population for NEA and thus any variation would have greater effect. This hypothesis is supported with the results from the distribution analyses (with the state average as well as the alternate averages) for both CA and TX as the NEA population incurred more significant differences than were observed for the other states. It should also be noted that the data set for California was not obtained as for the other states. The data were obtained through the National Center for Health Statistics rather than the California Department of Vital Statistics. As a result, the mortality data are for places of greater than 10,000 persons and does not include as many smaller communities as the data for other states. For this reason the 221 cities/towns in the California data set are also more urban than the communities for the other states.

One anomalous determination of significance occurred for Texas after recalculation with the alternate averages. This was for prostate cancer for NEAM aged 65-84 years. This could also have occurred due to the changes in Hispanic designations for the population. Additionally, it is possible this was a chance occurrence since each set of distribution comparisons for each state had a 0.05 probability of a significant result. The observations
for all six states could have resulted in a significant occurrence by chance. From Tippett’s formula in Equation 6 this probability may be as high as 25%.

6.2 Statistically Significant Individual Communities

Although the K-S method allows researchers to determine whether or not the set of communities is randomly distributed, it does not identify individual communities that may be significantly different by chance. In order to determine the significance of a community’s cancer rate it was necessary to evaluate the city or town’s individual binomial probability. Individual probabilities for five major adult and two major pediatric cancers for EAM and EAF were calculated. The significant results from this analysis were indicated on Table 2. Out of the nearly 27000 total trials of the 2216 communities, the seven cancers and age and gender cohorts, 506 were found to be significantly different. Of these 506, approximately 55% had been previously indicated with the Kolmogorov-Smirnov analysis of the distributions. Of the remaining significant trials, 80 had been identified by the K-S test prior to the recalculation with the alternate average. This shows an overwhelming congruence between for the distribution analysis and the individual community analysis, although there were a minority of significant results that were not detected with the Kolmogorov-Smirnov distribution analysis. These results also indicate that although the K-S test cannot detect small differences for a small number of communities from the perturbation tests, the concordant results with this second independent effort indicates the K-S test has succeeded in appropriately determining risk factors which vary throughout communities in the United States. Of the cancers and cohorts that were determined to be significant in a number of communities, lung cancer had by far the largest number. This was not unexpected as the distributions of lung cancer mortality rates were significantly different for several states for both genders and ethnic groups. For the communities that were significantly different for any of the cancers and cohorts, it is interesting to note that Woburn, Massachusetts, as well as the communities of Cape Cod, Massachusetts were not among them. Additionally analysis of Niagara Falls, New York, the site of Love Canal, showed that for none of the
cancers were its mortality rates significant except for lung cancer. These results are consistent with the results from Janerich’s study [24].

To determine the location of the significant clusters ArcView™, a mapping program was utilized. To map cities as points on a United States map a shapefile was obtained from the United States Geological Service (USGS), which allowed the majority of communities to be mapped in the data set. For some of the smaller communities, such as the burroughs in Pennsylvania the shapefiles did not provide their location. Figure 23 shows a map of the United States and the location of the communities that were significant for at least one of the seven cancers and one of the two cohorts studied. From the map in Figure 23 it is obvious that a good number of the significant communities are located in metropolitan areas; however, there were several that fell outside of urban areas. Because of the large number of significant communities for lung cancer these data were remapped exclusive of lung cancer and on a larger scale and are shown in Figure 24. In this figure it is more obvious how most of the significant communities occur in urban areas. These results were consistent with the distribution analyses that also suggested that urban areas influence the overall distribution of community cancer mortality rates.
Figure 23: Communities with at least one significant trial for seven major adult and two major pediatric cancers for EAM and EAF.

Significant communities were marked on the map with a large black dot.

Arrows point to major metropolitan areas with significant communities.
Figure 24: Communities with at least one significant trial for six major adult and two major pediatric cancers for EAM and EAF. (Lung cancer results were excluded)

Significant communities were marked on the map with a black dot.
6.3 Space-Time Considerations, Breast Cancer as an Example

The aggregation of the data as the sum of all death years as well as the analysis of individual communities does not address the chance that significant differences may occur in smaller time intervals or among groups of communities that may otherwise remain hidden in the larger analyses. In order to provide consideration for the possible clustering of community cancer mortality rates in space and time, three scenarios were evaluated. First, breast cancer mortality rates for the available communities were mapped utilizing ArcView™ software and shapefiles available from the United States Geological Survey. These maps can visually provide evidence for spatial clustering that may require future in-depth analysis. Second, breast cancer mortality rates were calculated in five-year intervals rather than the aggregate of all available death years for each state. These individual binominal probabilities for the number of deaths that occurred in each community were calculated for each time interval. The communities with one or more significant trials were recorded and again viewed with ArcView™ software to determine their locations. Last, an example for a space-time statistic was chosen (breast cancer for Massachusetts) because of previous reports of significant communities on Cape Cod, Massachusetts. The communities and time intervals were evaluated in both space and time, and the most likely clusters were recorded.

6.3.1 Consideration of Spatial Clustering

To consider the spatial distribution among communities mortality data were mapped for adult breast cancer (65-84 years) for EAF. The mortality rates were mapped using ArcView™ software and the USGS city shapefiles. With the maps it is possible to visually observe whether or not there is evidence of geographic clustering. The mortality rates for each city and town were plotted as a small circle with a grey organized as deciles. The lightest grey circles exhibited the lowest mortality rates and the black circles represent the communities with the highest mortality rates. White areas on the map represent areas between communities and communities for which data were unavailable.
The breast cancer mortality rate map is shown in Figure 25 with the area for MA, NY and PA enlarged for clarity. From this map it was possible to observe any areas that may show evidence of geographical aggregation. Outside of the metro area of Boston, Massachusetts the mortality rate data shown does not suggest geographical variation. This is investigated further with a temporal analysis in the next section, 6.3.2. From this map it is also clear that the available communities in California available for analysis were predominantly confined to the metropolitan areas of the state.
Figure 25: Spatial map of community specific mortality rates for breast cancer.
6.3.2 Consideration of Temporal Clustering

The data were further analyzed by dividing the death years available into five-year intervals. This was done for breast cancer data for the communities of all six states for EAF. For each of these time intervals the binomial probabilities for all communities based on the number of deaths, \( d_k \), from breast cancer for the cohort, the cohort population size, \( n_k \), and the state average, \( \rho \), were calculated again with Equation 1. These probabilities were recorded and compared with the 95% probability level adjusted by the Bonferroni inequality. This resulted in the same p-value as for the first individual community analysis (\( p < 1.13 \times 10^{-5} \)) as the basis for determining the probability limit was the number of communities for all six states, 2216.

The communities that were significantly different utilizing this probability level for any time interval were recorded for each interval. The location of these communities were again mapped using ArcView™. The significant communities for any of the time intervals were plotted as a grey dot in Figure 26. The dots overlayed with an X were those with increased mortality for the significant occurrence. Again it was observed that the majority of significant locales were in urban areas. There also were no significant occurrences for Woburn, Massachusetts, Niagara Falls, New York or for the towns on Cape Cod, Massachusetts.
Figure 26: Spatial map of communities with one or more significant time intervals for breast cancer among EAF.
6.3.3 Space-Time Scan Statistic of Massachusetts Breast Cancer

The previous look at the spatial and temporal distribution for breast cancer among the communities of the six states have indicated that the spatial distribution of breast cancer mortality rates appear to be random. Additional analysis of the binomial probabilities for five-year intervals for each community led to the conclusion that significant communities for one or more time periods were again confined to urban areas. To address questions of relationships that may arise in space and time, it is necessary to investigate the binomial probabilities for groupings of communities over different time intervals as well as groups of time intervals.

Breast cancer for Massachusetts was chosen for this analysis as significant community groupings for a particular time interval (1982-1990) have been indicated with studies of breast cancer incidence on Cape Cod, Massachusetts [33-34]. Therefore in order to investigate breast cancer in Massachusetts more robustly a space-time statistic was employed. In order to perform the spatio-temporal analyses, each time period and combination of time periods must be investigated. Likewise, each community grouping must also be investigated. This was possible by setting up a “moving window” which “visited” each time period starting with time $t_1$ and summing to $t_n$, time $t_2$ and again summing to $t_n$, ... to $t_n$. For ease of analysis the time periods were partitioned into three-year intervals. While this “moving window” covered the various combinations of time intervals, the window also incorporated the groupings of nearest towns.

For each of the time periods each community and its combinations of nearest neighbor communities were analyzed. For these analyses the nearest ten communities were determined. Like the moving window moves with time so it does for the nearest communities. First the community itself is analyzed for all of the possible time periods, explained above, then the data from the next nearest community is added and analyzed
again for the set of time periods until all nearest communities have been added to the first. This is done for every community in the set.

In order to determine the nearest communities, the latitude and longitude coordinates were obtained from the United States Geological Survey (USGS). These values were used to calculate the distance between any and all pairs of communities. The distances were calculated for each town combination, and the nearest ten communities to every community were selected for cluster analysis. Equation 11 below was used to calculate the distance between town pairs.

\[ d = \text{acos}((\text{sin}(\text{lat1}) \times \text{sin}(\text{lat2}) + \text{cos}(\text{lat1}) \times \text{cos}(\text{lat2}) \times \text{cos}(\text{long2-long1})) \times r) \]  
(Equation 11)

where,

- \( d \) is the distance between the two cities/towns
- \( \text{lat1} \) = the latitude of city1/town1 (in radians)
- \( \text{lat2} \) = the latitude of city2/town2 (in radians)
- \( \text{long1} \) = the longitude of city1/town1 (in radians)
- \( \text{long2} \) = the longitude of city2/town2 (in radians)
- \( r \) = the radius of the earth (3963 mile)

The analyses were performed as described for breast cancer (65-84 year olds) for the 251 available towns in Massachusetts. Of the nearly 125,000 trials for the various combinations of towns and time periods (45 different time interval combinations and 2761 community combinations) the only areas with significantly different breast cancer mortality rates (p<2 x 10^{-7}) were towns confined to the Boston metropolitan area.

Another combination of communities was significant utilizing the limit of 4 x 10^{-7}; however, this was not consistently observed for more than one time period and could still be explained by chance as this probability level is indicated with 95 % confidence limits and adjustment with the number of trials with the Bonferroni limit. The communities on
Cape Cod were again not significantly different even with the less stringent probability level of $4 \times 10^{-7}$.

None of the previous analyses have indicated significantly different breast cancer mortality rates for Cape Cod. To also directly compare the mortality data available for the towns of Cape Cod (with the exception of five towns that were too small and population data were not available) and the Massachusetts average, the age-specific breast cancer mortality rates were compared. The mortality data were summed for EAF in five-year age intervals, and the mortality rates were calculated with the corresponding population data. Figure 27, below, shows the composite breast cancer curve for Cape Cod versus Massachusetts. It is obvious there is no significant difference between the mortality rate curves in this case, even though Cape Cod has been touted as an area of dramatically higher breast cancer incidence. The two curves were not significantly different by the Log-Rank test ($p<0.05$).

Therefore, even with the claims of higher breast cancer risk on Cape Cod, the studies of breast cancer incidence [33-34] with three forms of space and time investigation do not support claims of geographical clustering of breast cancer in Massachusetts. Studies of breast cancer for Massachusetts through spatial mapping, time interval probabilities as well as a space-time statistic have shown that outside of urban aggregation, breast cancer mortality rates are distributed randomly. This again calls into question the use of incidence data for particular forms of cancer as well as studies with a limited number of communities. Although these results do not preclude the possibility that a community for a certain time period is significantly different due to some risk factor, these studies do show that chance variation is the driving force for community cancer risk; these analyses have also served to illustrate the extremely unlikely probability of finding any community which is significantly different.
Figure 27: Cape Cod versus Massachusetts age–specific breast cancer mortality rates, death years 1969-1995.
7. Conclusions

Historical and large-scale geographical studies have shown variations in cancer rates indicating the influence of risk factors, which may include genetic, medical, socioeconomic and/or environmental factors. However, the study of the distributions of community cancer mortality rates indicate carcinogenic risk at the community level are not significantly affected by these factors. In fact, the majority of the community distributions for the 21 most common adult and 6 most common pediatric cancers for three age- (0-19, 65-84 and ≥ 85 years) and four ethnicity/gender groups (EAM, EAF, NEAM and NEAF) for six of the largest states in the United States (California, Florida, Massachusetts, New York, Pennsylvania and Texas) are wholly distributed according to chance expectation. Of nearly 800 separate “chance” distributions which were compared with the corresponding observed distributions with the Kolmogorov-Smirnov test, 16 were found to have observed distributions significantly different from the expected by chance distributions, and 12 of the 16 were for lung cancer.

Lung cancer mortality rate distributions were significantly different for a number of states, age and ethnicity/gender cohorts. Due to the known influence of smoking on lung cancer risk [1-2] and evidence of geographic smoking variations that correlate with lung cancer rates [60], it is a valid assumption to attribute the differences among community lung cancer rates to the relative smoking prevalence among those communities. The other exceptions, such as prostate cancer in Texas for NEAM, prostate cancer in California for EAM and NEAM and breast cancer in California for EAF all aged 65-84 years, may be due to errors in the population estimates created from the delineation of Hispanic origin beginning with the 1980 census. Additionally, the data available for the state of California are more urban in nature as the communities represented are of a population size greater than 10,000 persons. This has also been illustrated with the breast cancer mortality rate maps.

The second approach to determining community risk with individual binomial probabilities for each community compared to stringent probability levels with the
Bonferroni inequality have been consistent with the results of the Kolmogorov-Smirnov distributions analyses. The cancer mortality rates for all 2216 communities in the set for all 6 states were explored for five major forms of adult and two forms of pediatric cancer for European American males and females for the age groups 65-84 years and 0-19 years, respectively. The binomial distribution was utilized to calculate the probabilities for the observed number of deaths for each cancer and cohort given the community’s cohort population size and the state average mortality rate in conjunction with the Bonferroni limit for the 95% confidence interval. Using this method, of the 26,592 trials for the communities, cancers and cohorts, 506 trials were found to be significantly different. Of these significant trials, nearly 55% had been identified with the Kolmogorov-Smirnov analyses. Of the remaining significant trials, 80 occurred for cancer/cohort groups that had indicated an urban/rural difference. Maps of these significant communities also indicated that significant difference occurred due to residence in urban locales.

Finally, studies of breast cancer in space and time also indicated an urban effect for “breast cancer”. With spatial maps, time interval probability analysis as well as a space-time statistic, no significant spatial and temporal aggregation was indicated outside of urban communities. These results support the primary conclusion that, outside of urban/rural differences, community cancer mortality rates are distributed by chance variation.
8. Future Research

It has been demonstrated that individual community cancer rate studies have proven to be a difficult if not futile endeavor. Additionally it has been demonstrated with a large set of community cancer mortality rates, that the majority of mortality rates at the community level are distributed wholly due to chance variation with few exceptions. Therefore, future study should focus on a larger scale where differences are less likely to be due to chance occurrence and risk factors can be measured with greater accuracy. These studies should focus on urban/rural, statewide or regional differences as well as worldwide cancer rate differences.

With the state data already compiled it would be possible to categorize the communities into urbanicity categories to further probe the differences that exist between urban and rural areas of the country. For instance, with the Massachusetts data set, each town was classified according to population density and general industrial categorization. These towns were grouped into three residential categories: urban, suburban, and rural through a combination of population density, industrial classification as well as proximity to the largest cities. The age-specific mortality rates were calculated for the residential groups with the summed death records from 1969-1995 and the corresponding population data. The age-specific lung cancer mortality rates were plotted for European American males (EAM) in Figure 28. The lung cancer mortality rates between the urban and rural areas of Massachusetts were found to be significantly different for this group (p <0.005, Log Rank test in conjunction with the Bonferroni inequality).

Additional analyses with the Massachusetts data set discovered significant differences for many major forms of cancer due to population density in addition to many other forms of death such as cardiovascular and cerebrovascular disease as well as respiratory diseases such as asthma and emphysema for both males and females and European and Non-Europeans. Many of these diseases have yet to be plausibly linked to environmental pollution and the reason for these differences are completely unknown. Figure 29 shows the urban and rural difference for breast cancer in Massachusetts.
Figure 28: Massachusetts age-specific lung cancer mortality rates for EAM, death years 1969-1995 according to population density.
Figure 29: Massachusetts age-specific lung cancer mortality rates for EAF, death years 1969-1995 according to population density.
Another area in need of further study, which is possible with mortality records is the regional or statewide differences that exist within the United States. A subset of the data compiled by the National Center of Health Statistics was available from the Harvard-MIT Data Center. This data set consists of all death records in the United States from 1968-1991. Distribution analyses (similar to the method used for community cancer rates) of these data are also possible. The distribution of cancer mortality rates across the 50 states and Washington D.C. were plotted and compared with the distribution expected by chance using the K-S test for colon and breast cancer and the respective national average mortality rates and the cohort populations for each state. Mortality rates for those aged 65 to 84 at the time of their deaths were calculated as an aggregate of all death years (1968-1991). Census Bureau data for each state and Washington, D.C. for the 1970, 1980 and 1990 census years are available on an age- and ethnicity- specific basis. With these data, it was found that the distributions of cancer mortality rates across the 50 states and the District of Columbia were very broad. When the observed distributions were compared to the expected, it was clear that the chance expectation cannot explain the actual distribution of cancer mortality rates across these 51 principalities in the United States. Figure 30 below show these distributions for EAM colon cancer and EAF breast cancer for those aged 65 to 84 years old, respectively. This confirms the idea that risk factors exist throughout the states in the country.

In addition to the investigations of urban and rural differences as well as the statewide differences in the United States worldwide variation should also be investigated. Figures 12 and 13 previously illustrated the large differences that exist between the United States and Japan. These large populations are clearly significantly different, and the reason(s) for these differences are not known. Studies of this magnitude not only clearly and unmistakably identify areas at significantly different risk, but also enable researchers to utilize data from these large populations to allow for more robust analysis of the factors which caused these differences.

In conclusion it is clear that there are risk factors, which cause significant differences in cancer rates between countries, and within subpopulations in the United States; however
it has been shown that at the community level these risk factors do not significantly influence cancer risk. Therefore, studies of individual communities will not further the aim of understanding the etiology of cancer in the general human population. Research to this end would be most successful if the focus were shifted from small, localized populations such as communities to large-scale populations such as worldwide, regional/state and urban/rural populations.
Figure 30: Distribution of colon and breast cancer mortality rates for EAM and EAF, respectively for the 50 States and the District of Columbia for the death years 1968-1991.

The statistical expectation is given as the solid line, and the observed mortality rates are represented by the black bars.
Literature Cited


[72] Marie Pees, United States Census Bureau, Division of Special Tabulations, personal communication.


Appendix

Mary Anne Freedman and Krasowski, Robert
Director, Division of Vital Statistics
National Center for Health Statistics
6525 Belcrest Road, Room 820
Hyattsville, MD 20782
(301) 458-4219
fax: (301) 458-40344

Rod Palmieri
State of California - Health and Human Services Agency
Center for Health Statistics
304 S. Street
P.O. Box 730241
Sacramento, CA 94244-0241

Gary Sammet
State of Florida
Department of Health
Office of Vital Statistics
P.O Box 210
Jacksonville, Florida 32231-0042
904-359-6900 ext.1053

Craig Edelman
Division of Health Statistics
Pennsylvania Department of Health
555 Walnut Street - 6th Floor
Harrisburg, PA 17101-1900
717-783-2548

Elaine Trudeau and Charlene Zion
Massachusetts Department of Public Health
Vital Registry
Boston, MA
617-753-8670

Phillip Quickenton, Gene Therriault and Robert Draiss
New York State Department of Health
Vital Records Section
IRB Research Unit
P.O. Box 2602
Albany, New York 12220-2602
518-474-5245
Dr. Steven Schwartz and Louise Berenson
New York City Department of Health
Office of Vital Statistics
125 Worth Street
New York, New York 10013 Box 7
212-788-4589

Leland Carmichael
Statistical Services Division
Bureau of Vital Statistics
Texas Department of Health
1100 West 49th Street
Austin, Texas 78756-3195
512-458-7509

Marie Pees
United States Census Bureau
Department of Special Tabulations
301-457-2447