MORTALITY ASSOCIATED WITH ARSENIC IN DRINKING WATER

by

Dr Virendra Kumar Bharti

A thesis submitted to

The Faculty of Graduate Studies and Research

in partial fulfilment of

the requirements for the degree of Master of Science

School of Mathematics and Statistics

Ottawa-Carleton Institute of Mathematics and Statistics

Carleton University

Ottawa, Ontario, Canada

Your file Votre référence
ISBN: 978-0-494-44116-9
Our file Notre référence
ISBN: 978-0-494-44116-9

This thesis is dedicated to my family

Grandparents: Shri Rajaram and Smt Samundara Devi Bharti

Parents: Shri Rudra Shanka and Smt Swarna Lata Bharti

Wife: Smt Vibha Bharti

Son and daughter: Shekharendra and Shivangi Bharti

I heartily love you all.

ACKNOWLEDGEMENTS

I thank my thesis supervisors Dr. Shirley Mills for her valuable support and guidance throughout this thesis. Dr. Mills, your encouraging remarks made this thesis project a very good learning experience. I greatly appreciate your mentorship. This thesis could not have been carried out with out the guidance and valuable suggestions of the committee members and I thank for that.

I take this opportunity to thank the faculty and staff specially Ms Valerie Daley of the Department of Mathematics and Statistics for their support.

I am thankful to my family members Mr. and Mrs. Luxmikant and Munni Bharti, Mr. and Mrs. Rakesh and Pramila Goswami, Mr. and Mrs. Surendra Kumar Bharti, Mr. and Mrs. Chandradev and Kusum Bharti, Late Mr. and Mrs. Ram Shankar Bharti, Mr. and Mrs. Uma Shankar Bharti, Mr. and Mrs. Dinesh Chandra Bharti, Mr. and Mrs. Ratnesh Kumar Bharti, Mr. and Mrs. Awadhesh Kumar Bharti, Mr. and Mrs. Awanish Kumar Bharti, Mr. and Mrs. Rajnish Kumar Bharti, Shivkumari didi, Shaila didi, Santa didi, Shashikala, Nirmala, Dhirendra, Alok, Archana, Alpana, Mona, Bholu, Sunil, Shobha, Saroj and other children.

Finally, I would like to thank all my Ottawa's uncles, aunts and friends Ravi-Manju, Radheyshyam-Sharada, Baliram-Shanti, Somesh-Anju, Jagdish-Krishna, Hari-Mina, Umesh-Chetana, Kishor-Namita, Dinesh-Monica, Ashraf-Farhat, Kamal-Nina, Raman-Neetu, Shahid-Shaila, Shahid-Sehmina, Sanjay-Vandana, Jai-Mamta, Amitabh-Sarita, Chandramauli-Kiran and others for their encouragement.

Abstract

Arsenic is a natural element found in the environment in organic and inorganic form. The inorganic form is much more toxic and is found in ground water, surface water and many foods. This form is responsible for many adverse health effects like cancer and cardiovascular and neurological effects.

The present study is based on the available mortality data on lung, bladder and liver cancer for the endemic region of Taiwan. The purpose of the study is to see the effect of various predictors like Gender, Cancer Type (Lung, Liver and Bladder), Arsenic Concentration (High, Medium and low), Age and Person Years on Mortality, using a model-based approach to analyze this data. Three major analytical techniques: Logistic Regression, Poisson Regression and Negative Binomial Regression are used to assess the effect of these explanatory variables or covariates on mortality, which is the response or dependent variable.

The major finding of the study is that either all or most of the categories of the predictors are significantly associated with the mortality. The logistic model finds Age as the only significant predictor ($p < .0001$) associated with the mortality. Models based on the actual frequency counts like Poisson Regression and Negative Binomial find all the five (5) predictors significantly associated with mortality.

Table of Contents

List of Tables

1. INTRODUCTION

Arsenic has long been associated with development in humans. It is a metal found in rock and mineral formation in the earth's crust (approximately 2–5 μg/l, [61]). It is the 20th most abundant element in the Earth's crust. A trace amount of arsenic is found in all living matter. Exposure can occur through food, smoking, water and air.

Arsenic may be found in water flowing through arsenic-rich rocks. The sources are diverse, as presented below:

1. Earth's crust
2. Introduced into the water through the dissolution of minerals and ores. The concentration in some area is elevated because of erosion from rocks.
3. Industrial effluent to water.
4. Combustion of fossil fuels is a source of arsenic in the environment through atmospheric deposition.
5. Abundance in seafood is very much less harmful and readily eliminated by the body.

The general amount of exposure to arsenic through food, smoking, water and air are shown in Table 1.1 [58].

Table 1.1: Intake of inorganic and organic arsenic compounds in the general population. [58]

Source	Inorganic Arsenic (μg/day, p)	Organic Arsenic (μg/day, p)
Food	5–20	5–1000
Smoking	1–20	–
Water	<1–20	–
Air	0.05	–

Drinking water is the primary route of exposure. Arsenic concentration generally varies between 1–2 μg/l in most natural waters. It results in an arsenic intake of 2–4 μg/day assuming the consumption of 2 litres of water per day [65]. This appears very low compared to food exposure but in some parts of the world the concentration varies up to 3200 μg/l, resulting in very high exposure. The importance of the various exposure routes depends on the actual concentration. The major exposure route would be food if the arsenic concentration is less than say 1 μg/l and there is no pollution of air due to industries. But in places like Bangladesh and India where the ground water arsenic concentration could be up to 3200 μg/l, the intake through drinking water will be much higher than through food. In Canada, naturally occurring arsenic is found in rocks in Nova Scotia, and leaches into water. It may get into water through industrial waste discharge or particles deposited in dust. The primary source of exposure to arsenic for most Canadians is food. The arsenic level in Canadian drinking water is generally less than 0.005 milligrams per litre or 0.005 parts per million (ppm). It is tasteless and odourless so it is hard to tell its presence in the water without lab testing.

There has been no study outside of Taiwan of a population exposed to arsenic in drinking water with adequate size and categories showing disease and progression. Residents of the area started using artesian well water in the second decade of the 19th century (Wu et al., 1989: [65]). The study suggested significant increase in mortality among workers at smelters. The most common source of human exposure to inorganic arsenic is through drinking water and in most cases by natural contamination. It is

absorbed by the body after someone swallows it, not by skin while bathing. Short term exposure for a few days or weeks can cause:

- abdominal pain, vomiting and diarrhoea,
- muscular cramping or pain
- weakness and skin rash or/and flushing,
- numbness,
- burning or tingling sensation or pain in hands and feet,
- thickening of the skin on the palms and soles, and/or
- loss of movement and sensory responses

The long term exposure for many years can cause:

- thickening and discoloration of the skin,
- nausea and diarrhoea,
- decreased production of blood cells,
- blood vessel damage,
- abnormal heart rhythm, and/or
- numbness in hands and feet

Arsenic is an established carcinogen and long term exposure (0.1 – 0.9 mg/l) is related with Blackfoot disease [50, 52]. An exposure of 1.5 mg/kg of body weight can lead to death. The most common symptoms of long term low level exposure are variations in skin pigments, hyperkeratosis and ulcerations. Previous studies have shown that chronic exposure to arsenic water in access of 400 microgram per litre (μg/l) is associated with the risk of cancer of skin, liver, bladder and lung. The US Environmental

Protection Agency (USEPA) has lowered the maximum contaminant level (MCL) for drinking water arsenic from 50 to 10 microgram per litre (μg/l) in 2001 based on international data, analysis and research. This recommendation is on hold now. The 10 parts per billion (ppb) became law in the European Union in January 2001. But even at 10 ppb, the risk of developing one of the cancers is increased by 3 in 1,000. There are claims about arsenic's ability to cause some cancers and cure others. Researchers have found that the compound arsenic trioxide was dramatically effective against acute promyelocytic leukaemia (APL) that had become resistant to other chemotherapy. One study supports the hypothesis that arsenic may act as a co-carcinogen – not directly causing cancer, but allowing other substances, such as cigarette smoke or ultraviolet light, to cause mutations in DNA more effectively [46].

The excess risk of getting skin cancer by gender for arsenic intake calculated by USEPA (US Environmental Protection Agency) [57] and Brown et al. [11] is presented in Table1.2.

Table 1.2: Lifetime excess risk of skin cancer

Author	Gender	Lifetime excess risk of skin cancer per 1 μg/kg, d
USEPA [57]	Male Female	$2.45 * 10^{-3}$ $1.05 * 10^{-3}$
Brown et al.[11]	Male Female	$1.3 * 10^{-3}$ $0.6 * 10^{-3}$

The lifetime excess risk of bladder cancer calculated by National Research Council (NRC) [2] and Smith et al. [47] for an American population consuming 2 litres

of water per day with an arsenic concentration of 50 μg / litre is presented in the following table.

Table 1.3: Lifetime excess risk of bladder cancer

Author	Lifetime excess risk of skin cancer per 1 μg/kg, d
NRC [2]	1/1000
Smith et al.[47]	7.4/1000

The Natural Resources Defense Council (NRDC) [43] in the U.S.A. analyzed data provided by 25 states. Their conservative estimate suggests that more than 34 million Americans drink tap water containing an average level of arsenic that poses unacceptable cancer risks. Their best estimate suggests as many as 56 million people of these 25 states are drinking water with unsafe levels of arsenic. The arsenic enters into the water supply either from natural deposits in the earth or agricultural and industrial pollution. Industries in the U.S.A. release thousands of pounds of arsenic in the environment. Table 1.4 below shows the lifetime risk of dying of cancer from arsenic in tap water.

Table 1.4: Lifetime risk of dying of cancer in tap water [43]

Arsenic level in tap water (in ppb)	Approximate total cancer risk (assuming 2 litres consumed/d)
0.5 ppb	1 in 10,000
1 ppb	1 in 5,000
3 ppb	1 in 1,667
4 ppb	1 in 1,250
5 ppb	1 in 1,000
10 ppb	1 in 500
20 ppb	1 in 250
25 ppb	1 in 200
50 ppb	1 in 100

The finding of a significant dose-response relationship between arsenic concentration and population in Taiwan was thought to be unique because of the genetic make-up of the study population or the low protein content of the traditional diet. Another study in Cordoba, Argentina also suggests a significant dose-response relationship. There the study population exposed to arsenic-contaminated drinking water had very different genetic heritage from Taiwan because of high protein in their customary diet. The water in the Argentina study was also different as it is contaminated only with arsenic, not with other substances as was the case with the Taiwan study [32].

There are many countries in the world where arsenic in drinking water is more than the recommended limit or guideline value of 0.01 mg/L. These are Argentina, Australia, Bangladesh, Chile, China, Hungary, India, Mexico, Peru, Thailand, and the United States of America. Some of the examples of adverse effects are:

1. The estimated number of people in Bangladesh in 1998 exposed to arsenic concentration above 0.05 mg/l is 28–35 million and the number of those exposed to more than 0.01 mg/l is 46–57 million.
2. The estimated number of people in West Bengal, India, the border province to Bangladesh, in 1997 actually using arsenic-rich water is more than 1 million for concentrations above 0.05 mg/L and is 1.3 million for concentrations above 0.01 mg/L.
3. The USEPA has estimated that 13 million of the US population are exposed to arsenic in drinking water at 0.01 mg/L.

The millions of hand-pumped tube wells installed in Bangladesh since the 1970s have led to 95% of the 130 million populations being dependent on underground water. The U.S.A. data and analysis is interesting because, although its arsenic concentration is quite low in comparison to Bangladesh, Taiwan and West Bengal, the assessment of risk to U.S.A. residents is made by extrapolating non-U.S.A. data. The NRC 2001 risk assessment and the USEPA's 2001 drinking water standard revision from 50 μg / litre to 10 μg / litre were based on the Morales et al. analysis [41] of Wu et al.'s data from Taiwan [65]. The arsenic concentration in U.S.A. data of 10 μg / litre has not demonstrated carcinogenicity.

The present study is based on mortality data on lung, bladder and liver cancer for the endemic region of Taiwan. The initial study done by Tseng et al. [52] has generated great interest not only in Taiwan but worldwide. The data is from a study of the population of 42 coastal villages in six (6) south-western townships where Blackfoot disease is endemic. The townships were Peimen, Hsuechia, Putai, Ichu, Yensui, and Hsiaying. The data is reported by Wu et. al. [65] and grouped into three (3) categories by arsenic concentration.

a. Less than 300 parts per billion (ppb),

b. 300−590 ppb, and

c. 600 ppb and over

The purpose of this thesis is to examine the effect of various predictors like Gender, Cancer Type (Lung, Liver and Bladder), Arsenic Concentration (High, Medium

and low), Age and Person Years on mortality. This study uses only three (3) major categories of cancers, not all as done by Wu et al. [65]. The purpose is to do an in-depth study using a *modelling* approach. In this approach, the magnitude of a relationship can be determined rather than just significant effects. So here, the significant or non-significant effects are measurable.

Three major analytical techniques given below are used to assess the effect of these independent variables or covariates on the response or dependent variable.

1. Logistic Regression Model.
2. Poisson Regression Model.
3. Negative Binomial Regression Model.

This study aims to throw more light on arsenic and its effect on human mortality by accounting for the magnitude of the relationship between exposure to arsenic concentration and cancer. The study outcome may help in making future decisions about permissible arsenic levels in water based on the magnitude of its relationship with mortality. It will also throw light in future research, planning and management of medical, biotatistical and model related studies in general and arsenic-in-water-related studies in particular.

2. LITERATURE REVIEW

There are an enormous number of publications and reports about arsenic and cancer from drinking water. Arsenic is a mineral found in nature and mainly transported in the environment by water. There are many documents available explaining its toxicity [29, 60 and 62]. The U.S.A. Environmental Protection Agency (EPA) classified arsenic as a class 'A' human carcinogenic, based primarily on epidemiologic evidence, and produced quantitative risk estimates for both ingestion and inhalation routes of exposure [54]. In 1975, the US adopted the interim standard as 50 μg / litre in response to the 1974 Safe Drinking Water Act. This has further been updated in 2001 [42].

The present study is based on secondary data from Wu et al. (1989) [65] so the literature review is much more focused and concentrated on the most relevant articles directly related to the study. The major work on this topic was done by the U.S.A. 'Environmental Protection Agency' called EPA. Other major contributors are Wu et al. (1989) [65], Chen et al. (1992) [17], Tseng et al. [50, 51] and National Research Council of the U.S.A. One preliminary study has focused on bladder cancer and examined model sensitivity [4].

Blackfoot disease is a vascular disorder confined to a limited area on the south-west coast of Taiwan [64]. This disease starts with coldness or numbness in one or more extremities and intermittent claudication and progresses to black discoloration, ulceration, and gangrene. Amputation of the distal part of the affected extremities is

common in the end stages of the disease [51]. Arsenic has been suggested as one of the most important determinants of Blackfoot disease [17, 18 and 19]. Blackfoot disease was related to artesian well drinking water in the endemic area of Taiwan [20]. Peripheral vascular disease has been found to be associated with the high arsenic concentration well water in Taiwan [4, 20, 33 and 34].

Wu et al. [65] did one of the leading and pioneering studies in 1989 on the south-western Coast of Taiwan Island, where Blackfoot disease [19] has been endemic. They found a significant relationship between arsenic levels in well water and bladder, kidney, skin, and lung cancer in both males and females. Its relation to prostate and liver cancer was significant for males. They did not find an association for cancers of the nasopharynx, esophagus, stomach, colon, and uterine cervix, and for leukaemia. The study also found associations between arsenic level and peripheral vascular diseases and cardiovascular diseases in a dose-response pattern. Wu et al. [65] recommended extensive epidemiologic follow-up and further examination of the effects of carciogenesis, atherogenesis, and their interrelation. A U.S.A. study found association between much lower levels of arsenic and bladder cancer for smokers [6]. This study points to some link between arsenic and cancer. One study found association between bladder cancer mortality and arsenic in drinking water in Argentina [32]. The high arsenic content of drinking water from wells in the Bell Ville region of Cordoba, Argentina was associated with increased incidence of clinical skin alteration. The cancer death rate in nine (9) highly exposed towns in the Cordoba region was 24% compared to

15% for the whole region between 1949 and 1959 [5]. Overall bladder cancer accounted for 2.9% deaths in Argentina in 1980.

Tseng and Tseng et al. [51, 52] reported association between high-arsenic artesian well water and skin cancer and Blackfoot disease spread to the south-west Coast of Taiwan Island. This study, referred to widely as the Taiwan study, has been used to set the standard for arsenic levels. In 1985 Chen et al. [18] reported very high mortality rates and further increase to be significantly associated with the use of high-arsenic artesian well water of the Taiwan region. Chen et al. in 1988 [20] again reported significantly higher age-adjusted mortality from various cancers among residents of the same area of Taiwan. They found a significant dose-response relation between the arsenic level and age-adjusted mortality for cancers of the bladder, kidney, skin, prostate, lung and liver. Bladder cancer has been found more common among all the associated cancers. This cancer has showed the steepest slope and highest relative risk. The Taiwan study has shown a dose-response relationship between arsenic and internal cancers categorized into three categories: low, medium and high [65]. The bladder cancer mortality rate ratio for men and women has been found as 12.1 and 25.1 per 1000 respectively [65] among medium exposure groups exposed to a weighted average arsenic level of 480 μg per litre. There is a further study and report on water regulations in 2001 [58]. There is increased risk of lung and possibly liver cancer for Moselle vintners who drank arsenic-contaminated wine [45].

The ingested or inhaled exposure to inorganic arsenic through medicinal, occupational and environmental exposure is a human carcinogen of skin and lung [33]. Ingestion is an established cause of skin cancer [6, 20 and 17]. The most common sources of ingested arsenic are drinking water, medications, Fowler's solution (a tonic containing potassium arsenic) and arsenic insecticides [6]. There are many epidemiological studies in West Germany [12, 30 and 31}, Argentina [3, 7, 8, 17, 32, 47 and 49], Mexico [14, 15], Chile [9, 66], and India [16] showing signs of positive association between arsenic content in drinking water and skin cancer. Skin cancer or its precursor lesions were found among psoriatic patients treated with arsenic [14]. The prevalence rate of cancer in the endemic area of Taiwan was as high as 10.6/1000. There was also an increase in bladder and lung cancer mortality in the region of Northern Chile [9, 66].

The estimate of potency index of developing lung cancer for an American male who is exposed to 1 μg /kg /day inorganic arsenic through inhalation is estimated by Chen and Chen to fall in the range 4.6×10^{-3} to 2.4×10^{-2} [23]. The above estimate is based on the data from smelter workers in Anaconda, Montana [10, 36 and 31] and in Tacoma, Washington. The potency index for an American male who is exposed to 1 μg/ kg /day inorganic arsenic through drinking water for a 76 year lifespan is estimated by US Environmental as 1.3×10^{-3} [11]. This estimate is based on the prevalence of skin cancer among residents in an endemic area of chronic arsenicism and an unexposed area [52].

There was increased risk of internal organ cancers associated with inorganic arsenic exposure through drinking water not only in the confined endemic area but in 314 precincts and townships of Taiwan [21]. Malignant neoplasm of internal organs in persons exposed to medicinal as well as vintners and smelters have been reported [31, 35 and 45]. Vascular disease other than the malignant neoplasm was associated with arsenic intake among West German vinedressers [30]. It was also noted among the inhabitants of Antofagasta, Chile [9]. A cohort study of patients treated with Fowler's solution found a threefold increase in bladder cancer [28]. Another study of arsenic-poisoned patients in Japan found elevated occurrence of urinary tract cancer [53].

3. MATERIALS AND METHODS

3.1. DATA

This thesis is based on mortality data on lung, bladder and liver for the endemic region of Taiwan. The initial study done by Tseng [51] has generated great interest not only in Taiwan but worldwide. The data is from a study of the population of 42 coastal villages in six (6) south western townships where Blackfoot disease is endemic. The townships were Peimen, Hsuechia, Putai, Ichu, Yensui, and Hsiaying [64]. The data is originally reported by Wu et. al [65] and grouped into three (3) categories by arsenic concentration (Appendix 1);

d. Less than 300 parts per billion (ppb),

e. 300–590 ppb, and

f. 600 ppb and over

The data is taken from the Statistical Issues article [2].

3.2. DATA ANALYSES

The data was analyzed using SAS 9.1.3 service pack 2 and SPSS 14.0 for Windows. Descriptive statistics were used to check out distribution, noise and any other discrepancy in the data. Three major analytical methods given below are used to measure the effect of predictors or independent variables or covariates (Gender, Cancer Type, Arsenic Concentration, Age and Person Years) on response (dependent variable) which is death.

1. Logistic Regression Model.
2. Poisson Regression Model.
3. Negative Binomial Regression Model.

All the details and steps involved in the analysis are elaborated below under respective analysis headings.

3.3. LOGISTIC REGRESSION ANALYSIS

SAS software provides the option of logistic regression modeling called logistic procedure. The response (outcome or dependent) variable is death (Yes/No) and predictor (independent) variables are Gender (Male/Female/), Cancer (Lung, Liver and Bladder), Arsenic Concentration (High, Medium and low), Age and Person Years (continuous). Descriptive statistics were used to examine frequency distribution and measures of central tendency of categorical and continuous variables respectively. Categories were collapsed to remove singularities or zero cells.

The mathematical model of the logistic regression begins with the explanation of the logistic function:

$$f(z) = 1/(1+e^{-z})$$

The variable z represents the exposure to the set of risk factors i.e. Gender, Age, Cancer Type, Arsenic Concentration and Person Years. The function f(z) represents the probability of death, given the set of risk factors mentioned above. The variable z is a measure of the total contribution of all the risk factors used in the model and is known as the logit. The variable z is defined as:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

where β_0 is the "intercept" and β_1, β_2, β_3, and so on are the "regression coefficients" of x_1, x_2, x_3 respectively. A positive coefficient means the factor increases the risk of death and vice versa for a negative coefficient.

3.3.1. UNIVARIATE (SIMPLE) LOGISTIC REGRESSION ANALYSIS

Descriptive statistics have been obtained to examine extremities, data distribution and other details. Special attention has been paid to zero cell frequencies. Categories have been collapsed to handle zero cell problems. Smoothed scatter plots are applied for continuous predictors to ascertain importance of the variable, possible presence and effect of extreme observations and appropriate scale. SAS software provides the option for plotting the smoothed values on the logit scale, which make decision about the possible scale easier.

The significance of the independent variable was assessed by comparing the value of −2 log likelihood ratio (called 'D') with and without the independent variable in the equation. D has the following expression [31]:

$$D = -2\ln\left[\frac{\text{(likelyhood of the fitted model)}}{\text{(likelihood of the saturated model)}}\right]$$

The change in D due to the inclusion of the independent variable in the model is obtained as:

$$G = -2\ln\left[\frac{\text{(likelihood without the variable)}}{\text{(likelihood with the variable)}}\right]$$

The above test for univariate logistic regression is equivalent to Pearson's chi-square for categorical variables and the two sample t-test for continuous variables. These tests were undertaken to examine relationship between predictor and outcome variable.

Under the hypothesis β_1 equal to zero, the statistic G follows a chi-square distribution with 1 degree of freedom. Additional mathematical assumptions are also needed but for the present study they are rather non-restrictive and involve having a sufficiently large sample size.

The calculation of the −2 log likelihood and the likelihood ratio test are standard features of the SAS software. This simplifies checking for the significance of the addition of a new predictor to the model. In case of a univariate logistic regression (single independent variable), first the model has been fitted with the constant term. Then the model containing the independent predictor has been fitted. This gives rise to log

likelihood. The likelihood ratio test G is obtained by taking the difference between these two values and is used to test the significance of the omitted independent variable.

Two other similar statistical tests provided by SAS software have also been used. These are the Wald and Score tests. The Wald test is obtained by comparing the maximum likelihood of the slope parameter, $\hat{\beta}_1$, to an estimate of its asymptotic standard error. The equation of the statistic routinely printed by SAS software is presented below.

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Both the likelihood ratio test G, and the Wald test, W, require the computation of the maximum likelihood estimate for β_1.

There is another test called the Score test, which does not require the above computation. This is also routinely printed by SAS software in the output of the logistics regression. The Score test is based on the distribution theory for the derivatives of the log likelihood. The test statistic for the Score test (ST) is

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{y})^2}}$$

The methods for testing coefficients of the logistic regression are similar to those of linear regression; however we use the likelihood function for a dichotomous variable such as we have in the present study.

Initially it was thought to screen in all the predictors with $p<0.25$ for the multivariate logistic regression analysis but since all the predictors are very important to examine, it was decided to use all of them at the multivariate stage regardless of p value. This decision was taken in light of possible confounding effects of the predictors.

3.3.2. MULTIVARIATE LOGISTIC REGRESSION ANALYSIS

Collinearity among the predictors was assessed by the correlation matrix and verified by standard error. A huge value of standard error indicates problem (s) with the data including collinearity. The predictors having lesser relationship (based on p value) with response (death) are dropped from the model.

The estimation method for multivariate logistic regression analysis is the same as for simple logistic (one variable), which is maximum likelihood. The first step after the multivariate model fitting is to assess the significance of the predictors. The −2 log likelihood ratio test for overall significance of the 5 predictor (independent) variables

Gender, Cancer, Arsenic Concentration, Age and Person Years in the model is performed in exactly the same manner as in the univariate (single independent variable) case.

The test is based on the statistic G given in 3.3.1. The only difference is that the fitted values under the model are based on the vector containing 5+1 parameters (5 predictors + 1 constant). Under the null hypothesis, the five (5) slope coefficients for the covariates in the model are equal to zero. Since our goal is to obtain the best fitting model while minimizing the number of parameters, the logical step is to fit the reduced model containing only the significant variables. But this step is done by dropping the most non-significant predictors one by one, not all at once. At every step the reduced model is compared with the previous model. The difference between the two models is due to the exclusion of the non-significant predictor.

The multivariate analog of the Wald test provided by SAS Statistics software is obtained from the following vector-matrix calculation:

$$W = \hat{\beta}'[\hat{Var}(\hat{\beta})]^{-1}\hat{\beta}$$

and distributed as chi-square with 5+1 degrees of freedom under the hypothesis that each of the 5+1 coefficients is equal to zero in the case where all the predictors are fitted against response.

The multivariate analog of the Score test for the significance of the model is based on the distribution of the five (5) derivatives of L (β) with respect to β. This is provided by SAS Statistics output.

The SAS Stat software package for logistic regression routinely calculates confidence interval estimates for coefficients, thus eliminating the computational burden.

The odds ratio (OR) is a measure of association widely used in health and epidemiological studies. It approximates how much more likely is the outcome (death) with predictor in one category as compared to predictor in another category. The odds ratio is our parameter of interest due to its ease of interpretation. The simple relationship between a coefficient in the regression model and the odds ratio ($OR = e^{\beta_i}$) is the fundamental reason for logistic regression being so powerful. Again, the SAS Stat software package for logistic regression routinely calculates the odds ratio and its confidence interval.

A confounding effect of a covariate is ascertained by comparing the estimated coefficient for the risk factor predictor(s) from models containing and not containing the covariate in question as a confounder. Any important change in the estimated coefficient for the risk factor suggests the covariate is a confounder and has been included in the model, regardless of significance level of its coefficient. The modification effect of a predictor is ascertained by adding interaction terms to the model. If a predictor is an

effect modifier as well as a confounder then its confounding effect is of secondary importance and hence only the interaction term is included for that predictor [1].

To summarize, here is what has been done, presented in point form for model building:

1. After fitting the multivariate model, the importance of each variable has been verified by the Wald statistic and comparison of estimated coefficients with the coefficient of the model containing only that variable has been made.
2. Predictors that do not contribute to the model based on the above criteria are dropped from the model and a new model without that predictor is re-fitted.
3. The new model is compared to the old model using the likelihood ratio test.
4. If the coefficients have changed significantly then the dropped predictor is again included in the model, since this indicates the dropped predictor(s) is providing a needed adjustment to the effect of the predictors that remained in the model.
5. This process of dropping, refitting and re-inclusion is continued until the final model has been figured out.
6. All the left-out variables are added back one by one to this model to see if they contribute significantly to the likelihood ratio. If they do, then they are included in the model.
7. Linearity of the continuous predictor(s) is (are) checked. This test is provided in the SAS software by the GAM procedure.

8. Once the final model is chosen, the interaction terms are added one at a time to the model. They remain in the model if significant; otherwise they are dropped.
9. Standard errors, coefficients and odds ratios are verified for any possible problem.
10. This is the final Multivariate Logistic Model.

3.4. POISSON REGRESSION ANALYSIS

In SAS, the GENMOD procedure fits the Poisson regression model. Here the response (number of deaths) has been used as a count in the original form instead of grouping into two categories (No death, Death) as was the case for logistic regression modeling. Predictors are considered in the same form as in logistic regression. The reason for doing Poisson Regression Modeling was non-inclusion of all the important predictors (Gender, Cancer Type, Arsenic Concentration and Person Years) except one (Age) in the final logistic regression model. We wished to determine if these variables might prove significant in modelling the *number* of deaths.

In Poisson regression it is assumed that the response variable (Y), number of deaths, has a Poisson distribution given the predictors X1, X2, …, Xn,

$$P(Y = k \mid x_1, x_2, \ldots, x_n) = e^{-\mu}\mu^k/k!, \qquad k = 0, 1, 2, \ldots,$$

where the log of the mean μ is assumed to be a linear function of the response variable death. That is,

$$\log(\mu) = \text{intercept} + b1*X1 + b2*X2 + \ldots + bn*Xn,$$

which implies that μ is an exponential function of predictor variables,

$$\mu = \exp(\text{intercept} + b1*X1 + b2*X2 + \ldots + bn*Xn)$$

The maximum likelihood method is again used to estimate the parameters of the Poisson regression model. Adequacy of the model begins with the assessment of the descriptive statistics of the response count data. If the mean and variance are very different then the model is likely to be over- or under-dispersed. The deviance and Pearson chi square values reflect the fit of the data to a Poisson distribution in the regression. The deviance function is given below:

$$Deviance = 2\sum_{i=1}^{n} y_i \ln\left[\frac{y_i}{\hat{\mu}_i}\right] - (y_i - \hat{\mu}_i)$$

Deviance and Pearson Chi-Square divided by the degrees of freedom are used to determine over- or under-dispersion. For a Poisson distribution, the mean and the variance are equal, which implies that the deviance and the Pearson statistic divided by the degrees of freedom should be approximately one. Values greater than one indicate over-dispersion, that is, the true variance is bigger than the mean; values smaller than one indicate under-dispersion, where the true variance is smaller than the mean. Evidence of over- or under-dispersion indicates inadequate fit of the Poison model. Corrective measures in case of over-dispersion include using a Negative Binomial Regression Model instead of a Poisson Regression Model.

Significance of parameters is assessed based on chi-square and Wald 95% confidence limits. Testing of sets of model terms are conducted using likelihood ratio tests. All the above tests and statistics are provided by SAS software.

Over-dispersion is tested with a likelihood ratio test based on the Poisson and the Negative Binomial distributions. This procedure tests equality of the mean and the variance (imposed by the Poisson distribution) against the alternative that the variance exceeds the mean. The variance = mean + k mean^2 (k >= 0; the negative binomial distribution reduces to Poisson when k = 0) for the negative binomial distribution.
The null hypothesis is:

H_0: k = 0

and the alternative hypothesis is:

H_1: k > 0

The following steps are carried out for the test:

1. Record the Log Likelihood (LL) value after running the Negative Binomial Regression Model.
2. Record the Log Likelihood (LL) value of the Poisson Regression Model.
3. Compute Likelihood Ratio (LR) statistic using LR test as: −2[LL(Poisson) – LL(Negative Binomial)]. The asymptotic distribution of the LR statistic has probability mass function of one half at zero and one half – Chi-sq distribution with 1 degree of freedom [11].
4. To test the null hypothesis (H_0): Reject H_0 if LR statistic > $\chi^2_{(1-2\alpha,\ 1\ df)}$.

3.5. NEGATIVE BINOMIAL REGRESSION ANALYSIS

In SAS, the GENMOD procedure fits the Negative Binomial Regression Model. The procedure is very similar to fitting of the Poisson Regression Model. The Negative Binomial distribution adds a quadratic term to the variance representing over-dispersion. It takes the form:

$$P(Y = k \mid x_1, x_2, \ldots, x_n) = \frac{\Gamma(k + \frac{1}{o})}{k!\Gamma(\frac{1}{o})} = (\frac{o\mu}{1+o\mu})^k (\frac{1}{1+o\mu})^{1/o}$$

where o is the over dispersion parameter and the variance is

$$\mu + o(\mu)^2$$

Here the log of the mean, μ, is a linear function of predictor variables similar tothe case of the Poisson Regression Model.

$$\log(\mu) = \text{intercept} + b1*X1 + b2*X2 + \ldots + bn*Xn,$$

which implies that μ is an exponential function of the predictor variables, i.e.,

$$\mu = \exp(\text{intercept} + b1*X1 + b2*X2 + \ldots + bn*Xn)$$

Instead of assuming Y (the number of occurrences of an event) follows a Poisson distribution, it is now assumed that it follows a negative binomial distribution. This basically relaxes the assumption about equality of mean and variance, which is a strict property of the Poisson distribution. The variance of the negative binomial is equal to $\mu + k\mu^2$, where $k \geq 0$ is a dispersion parameter.

The maximum likelihood method is used to estimate k as well as the parameter of the regression model for dependent variable log (μ). The criterion for assessment of goodness of fit and interpretation of the results is the same as in the case of the Poisson Regression Model.

4. RESULTS AND DISCUSSION

4.1. Data and Distribution:

The secondary data taken from the literature [65] is presented in Appendix 1. The basic characteristics of events (cancer death of 1 or more vs. no death) with respect to different independent or response variables are summarized in Table 4.1.2. There were a total of 234 cases.

Table 4.1.1: Characteristics of events and predictors

Characteristics	No. of Cases (%)		
	No death	Death (1 or more)	Total
Gender			
Female	39 (16.67)	78 (33.33)	117 (50.00)
Male	32 (13.68)	85 (36.32)	117 (50.00)
Age group (years)			
20–35	47 (20.09)	7 (2.99)	54 (23.08)
35–40	10 (4.27)	8 (3.42)	18 (7.69)
40–50	7 (2.99)	29 (12.39)	36 (15.38)
>50	7 (2.99)	119 (50.85)	126 (53.85)
Cancer Type			
Bladder	28 (11.97)	50 (21.37)	78 (33.33)
Liver	22 (9.40)	56 (23.93)	78 (33.33)
Lung	21 (8.97)	57 (24.36)	78 (33.33)
Arsenic Concentration			
Low (0–300)	21 (8.97)	57 (24.36)	78 (33.33)
Medium (300–600)	25 (10.68)	53 (22.65)	78 (33.33)
High (>600)	25 (10.68)	53 (22.65)	78 (33.33)
Person Years, continuous			
Cases	71	163	234
Mean ± SD	10201.07±8176.40	4811.52±4124.67	6446.81±6170.83
Person Years, quartiles			
Q1	7 (2.99)	50 (21.37)	57 (24.36)
Q2	10 (4.27)	50 (21.37)	60 (25.64)
Q3	21 (8.97)	39 (16.67)	60 (25.64)
Q4	33 (14.10)	24 (10.26)	57 (24.36)

The above Table 4.1.1 is derived from the table presented in Appendix 1. The cross table frequency totals of different sub-categories of predictors (Gender, Age Group, Cancer Type and Arsenic Concentration) and response (1 or more Death and No Death) are the same except for Age Group. The Age Group total frequencies are different because the age categories are collapsed to remove the zero or small cell counts, a

requirement for logistic regression to optimize the probability estimates. The cell counts of different subcategories within predictors suggest that there is no big difference between / among subcategories except for Age Group. This gives an *a priori* idea about the possible outcome of the logistic regression analysis.

The predictors (except person-years) are categorical as evident in Table 4.1.1. The summary statistics pertains to frequency and percent distribution for categorical predictors, and sample size, mean and standard deviation (SD) for continuous predictors. All these statistics are presented in the standard form of the cross table of response vs. predictors. The predictor Person-Year is a combination of number of people and their Ages. We have Age as a separate predictor in the model so an expected issue in the model could be collinearity, interaction and/or confounding.

4.2. LOGISTIC REGRESSION:

The logistic regression model is fitted and evaluated in four steps in order to reach the final model.

1. Diagnostic Testing
2. Bivariate Logistic Regression
3. Multivariate Logistic Model
4. Collinearity, Confounding and interaction testing
5. Final Logistic Regression Model

4.2.1. Diagnostic Testing:

The cell counts for all the categorical variables are ≥ 5 as presented in Table 4.1.1. The summary statistics with sample size for continuous variable Person Year looks good, with sufficient cell counts in both categories of Death (1 or more) and No Death. The standard deviation is at the higher end but the values of Person Year are also large and this justifies it.

The details of statistics, significance, validation and test outcomes related to some of the diagnostics are presented either in the bivariate or multivariate section. All the important variables are included in the final model. Though the criteria of $p<0.25$ for inclusion in the final model is met only by Age and Person Years other predictors are important from health and research perspectives and hence are included in the final model. There is not much control over including more predictors other than the ones provided by the data since the study is based on secondary data. The final multivariate logistic model does not include extraneous variables, only the significant ones. Again there was not much control over this but the best predictors were chosen from the available five. The independence of observations is verified from literature review [41, 64].

The standard errors give a good idea about the validity and any possible problems with the data and distribution of the bivariate and multivariate logistic models. Any big

standard error value points to a data or other problem. None of the standard errors for bivariate and multivariate logistic models are large. They all are around one, which is good. The details of these values are presented in the respective sections of bivariate logistic models (4.2.2.1).

4.2.2. Bivariate Logistic Regression Model:

The bivariate logistic model has been fitted after collapsing some of the lower cell frequency groups to the nearest one to remove singularities or zero cell. One of the continuous predictors (namely Person Year) has been tested for linearity. A smooth LOWESS plot was also applied to verify best categories for continuous variable.

The minimum and maximum standard errors among all the bivariate logistic regressions were found to be as 0.000032 and 0.624 respectively. The minimum was associated with continuous Person Year variables whereas the maximum was with the 35−40 age group categories. This indicates that there is no problem whatsoever with the data.

Detailed statistics pertaining to the bivariate logistic regression model are presented in Table 4.2.2.1. The response variable is no death (number of death = 0) and death (number of deaths > 0). The event variable probability to be modelled by the logistic model is number of deaths > 0.

Table 4.2.2.1: Results of the bivariate logistic regression analysis (N = 234)

Predictor	Type	Cat.	−2 LogL	χ^{2a}	df	p^1	Est.	p^2	SE[b]	OR[c]	CL[d]
Gender	Categorical		286.24	0.99	1	0.32					
		Female								1	
		Male					0.284	0.320	0.285	1.33	0.76, 2.32
Age	Categorical		155.92	77.9	3	<.0001					
		20−35		8						1	
		35−40					1.681	0.007	0.62	5.37	1.58, 18.24
		40−50					3.326	<.0001	0.58	27.82	8.85, 87.44
		>50					4.737	<.0001	0.56	114.12	37.96,343.1
Cancer Type	Categorical		285.51		2	0.42					
		Bladder		1.73						1	
		Liver					0.354	0.304	0.34	1.42	0.72, 2.80
		Lung					0.419	0.228	0.34	1.52	0.77, 3.01
Arsenic Concen tration	Categorical		286.57		2	0.72					
		Low Medium		0.65			0.247	0.483	0.35	1.28	0.64, 2.55
		High					9.9E-17	1.000	0.34	1.00	0.51, 1.96
										1	
Person Yr	Continuous		248.71		1	<.0001	-0.0002	<.0001	.000	1.00	1.00, 1.00
Person Yr	Quartiles	Q1	251.82	27.0	3	<.0001					
		Q2		4			2.284	<.0001	0.48	9.82	3.79, 25.38
		Q3		31.2			1.928	<.0001	0.44	6.87	2.91, 16.23
		Q4		3			0.937	0.014	0.38	2.55	1.21, 5.39
										1	

Cat.=Category, Est.=Estimate, a - Wald Chi-square, b - Standard Error, c - Odds Ratio,
d - Confidence Limits, 1 – Overall p value, 2 – Estimated Coefficients p value

The Age and Person Year variables were found highly significant ($p < 0.0001$). Other variables or predictors viz. Gender, Cancer Type and Arsenic Concentration were non-significant. This is a bit surprising and may require further investigation, especially Arsenic Concentration. One possible explanation could be the missing information from the available data. The predictor Age is also modelled with 6 age categories. The categories were 20−35, 35−40, 40−45, 45−50, 50−80 and 80−85. This has also been found highly significant with $p < 0.0001$. The detail of frequency and percent distribution is not given here since four (4) Age categories were used in the final multivariate logistic model. The screening criteria for predictors to be included in the multivariate logistic model are $p < 0.25$ or less. All the non-significant predictors did not meet this criteria but

considering their health, medical and biostatistical importance they have been screened in. The standard error for all the bivariate models is less than one. This suggests no issue with the data, cell counts or probability convergence. The odds ratios (ORs) for non-significant and significant predictors were non significant and significant respectively.

The linearity assumption for the Person Year continuous variable was tested using GAM procedure of SAS STAT software. The t statistics for Linear (Person Year) is found to be −6.72 with $p < 0.0001$. This suggests a significant linear relationship of Person Year with response being 'number of deaths'.

The categorical form of Person Year was tried since it makes the interpretation of odds ratio easier. The categories were made by quartiles. The value of 1 in the odds ratio column indicates the comparison category; i.e. this is the category with which other categories of the predictor were compared.

Confidence limits not including 1 suggest a significant odds ratio. Two significant odds ratios were found for Age and Person Year. Those who are in the age group of 35−40, 40−50 and > 50 were 5.4, 27.8 and 114.1 times more likely to have 1 or more deaths, respectively, than those in the age group of 20−35. This suggests that the groups of older people have greater likelihood of more deaths compared to the group of younger ones. Those who are in the first, second and third quartiles of Person Year were 9.8, 6.9 and 2.6 times more likely to have 1 or more deaths, respectively, than those in the fourth quartile. This suggests that as the Person Year increases the likelihood of 1 or more

deaths decreases. One possible explanation could be that there are few survivors in the higher age groups. So though the age of a group is less, there are more people in that group thus explaining why the Person Year of that group is larger compared to the older age group. It has already been established that the older age groups are more likely to have one or more deaths.

4.2.3. Multivariate Logistic Regression Model:

All the predictors were included in the multivariate logistic model, though three out of five did not meet the screening criteria of $p < 0.25$. But as explained earlier, considering the importance of these predictors, they were included in the multivariate model. The first three attempts to build a multivariate logistic model detect quasi complete separation of data points. A quasi complete separation occurs when values of the target variable overlap or are tied at a single or only a few values of a predictor variable. The symptoms are extremely large calculated values for the odds ratio point estimates or large standard errors. The analysis may also fail to converge. If quasi-complete separation is detected, the predictor variable(s) showing separation should be removed from the analysis.

Six (6) Age categories and Person Years based on quartiles were used for the first three multivariate logistic model attempts. The predictors showing quasi complete separation or least significance were dropped one by one during each attempt of building the first three models. The first, second and third predictors dropped from the model were

Arsenic Level, Cancer Type and Sex respectively. But the separation problem persisted. So it was decided to collapse the six age categories into four and refit the multivariate logistic model.

The first attempt of at this second series of multivariate logistic regression models consisted of using all the five predictors with Age as four categories and Person Year as four quartile categories. But the problem of quasi complete separation persisted. The predictors posing problems were Age, Person Year and Arsenic Level as evident from huge standard errors and odds ratio point estimates. So it was decided to look into the interrelationships of these variables. The Pearson correlation coefficient is calculated to observe the interrelationship between the response and predictors, and within the predictors. The correlation matrix is presented below in Table 4.2.3.1.

Table 4.2.3.1: Results of the Pearson correlation coefficient (N = 234) with p value

Variables	*Death*	*Age*	*Person Year*	*Gender*	*Cancer Type*	*Arsenic Level*
Death	1.000	0.699 (p<.0001)	−0.402 (p<.0001)	0.065 (p=0.322)	0.079 (p=0.225	−0.046 (p=0.488)
Age	0.699 (p<.0001)	1.000	−0.588 (p<.0001)	0.000	0.000	0.000
Person Year	−0.402 (p<.0001)	−0.588 (p<.0001)	1.000	0.049 (p=0.447)	0.000	−0.520 (p<.0001)
Gender	0.065 (p=0.322)	0.000	0.049 (p=0.447)	1.000	0.000	0.000
Cancer Type	0.079 (p=0.225)	0.000	0.000	0.000	1.000	0.000
Arsenic Level	−0.046 (p = 0.488)	0.000	−0.520 (p < 0.0001)	0.000	0.000	1.000

The correlation matrix presented above suggests that Age and Person Year are strong confounders since they have high significant correlation between themselves and at the same time with the response death. It is known that Person Year is a derived

variable from Age and Persons so the high correlation between Age and Person Year is not surprising. The high correlation between Person Year and response 'death' throws light on the unobserved variable 'persons'. Here Age and Person Year have significant positive and negative relation respectively with the response death. Since the variable Person Year is derived from Age and person, and Age is positively related to death, hence the unknown variable 'person' is negatively associated with the response 'death'. Other than this information, which is already ascertained, there is no extra benefit of including the Person Year variable in the multivariate model. This may pose a problem of over-fitting the model. The other reason for non inclusion of Person Year is the relative comparison of its statistics with the Age variable. The statistics -2 log likelihood (Age = 155.92, Person Year = 248.71 and 251.82 respectively for continuous and quartile form) and chi-square (Age=77.98, Person Year = 27.04 and 31.23 respectively for continuous and quartile form) of Age are statistically better than Person Year. The predictor Person Year is dropped from the model considering all these facts based on the analysis. So, the only significant predictor left in the multivariate model is Age, since Person Year is dropped and other predictors are non significant.

4.2.4. Collinearity, Confounding and interaction testing:

The correlation matrix displays the bivariate collinearity. The collinearity may be multiple involving more than two predictors. This is not an issue anymore since only Age is left in the model. The confounding effect of all the left-out predictors other than the Person Year is tested by adjusting the Age model with these predictors. Person Year is

not tried for the reasons discussed earlier in section 4.2.3. Interaction effects were tried first since they take precedence over confounding.

All the interaction effects of Age and other predictors were tried one by one in order of their significance level in the bivariate logistic regression analysis. The sequence was Gender (p = 0.32), Cancer Type (p = 0.42) and Arsenic Concentration (p = 0.72). The results of all the interaction effects are presented in the following Table 4.2.4.1.

Table 4.2.4.1: Results of the interaction effect of Age with other predictors

Variables	***−2LogL***	***P value[a]***
Only Age	155.92	<.0001
Age with Age*Gender	154.130	
Age		0.004
*Age*Gender*		0.65
Age with Age*Cancer Type	147.813	
Age		0.0002
*Age*Cancer Type*		0.137
Age with Age*Arsenic Concentration	151.423	
Age		<.0001
*Age** Arsenic Concentration		0.294

a - Wald chi-square p value from Type 3 Analysis of effects.

None of the interactions were found significant as is evident in the above table. Though it was interesting to observe the interaction effect of Age*Gender, since Gender was non-significant as determined earlier, nothing unexpected occurred. It may require more research or/and data to investigate it further.

The confounding effects were tested by adjusting the model with each predictor in order of their significance level and comparing the −2 log likelihood estimates before and

after the adjustments. The difference in −2 log likelihood is tested against the chi-square table value with 1 degree of freedom. The results are presented below in Table 4.2.4.2.

Table 4.2.4.2: Results of the confounding effect of predictors on Age

Variables	***−2LogL***	***Differences[a]***	***Adjusted effect***
Only Age	155.92		
Age adjusted by Gender	153.75	2.17	Non significant
Age adjusted by Cancer Type	152.15	3.77	Non significant
Age adjusted by Arsenic Concentration	154.49	1.43	Non significant

a – Differences between −2LogL of only age and Age adjusted by corresponding predictor

The chi-square table value at 5% level of significance with 1 degree of freedom for difference comparison is 3.84. The above table shows all the −2 log likelihood differences are less than 3.84. Thus it has been determined that none of the adjustment or confounding effects is significant. The adjustment effect by Cancer Type is however very close to being significant.

4.2.5. Final Logistic Regression Model

The final logistic model after all the assumptions, data noise, collinearity, confounding and interaction tests and validations has been found to have one predictor namely Age. This means it is the same model as the one predictor bivariate model. The form of the categories, frequency distributions, −2 log likelihoods, estimated coefficients, standard errors and p values related to this model is presented in Table 4.2.2.1. The results are discussed after Table 4.2.2.1. Most of the statistics details are provided in that table. Some other details of the final model are presented below.

Table 4.2.5.1: Results of the final logistic regression model (N=234)

Predictor	Category	No deaths (n)	>=1 deaths (n)	Intercept	Coeff	p value	OR
Age						<.0001	
				-1.904		<.0001	
	20–35	47	7				1
	35–40	10	8		1.681	0.007	5.37
	40–50	7	29		3.326	<.0001	27.82
	>50	7	119		4.737	<.0001	114.12

The findings based on Odds Ratio estimates suggest that:

1. Those who are in the age group of 35–40 were 5.4 times more likely to have 1 or more deaths respectively, than those in the age group of 20–35.
2. Those who are in the age group of 40–50 were 27.8 times more likely to have 1 or more deaths respectively, than those in the age group of 20–35.
3. Those who are in the age group of >50 were 114.1 times more likely to have 1 or more deaths respectively, than those in the age group of 20–35.

The mathematical form of the final logistic regression model with mortality as dependent variable and Age as independent variable is:

$$\text{Probability of 1 or more deaths} = [f(z)] = 1/(1+e^{-z})$$

where $z = -1.094 + 1.681*\text{Age}(35\text{-}40) + 3.326*\text{Age}(40\text{-}50) + 4.737*\text{Age}(>50)$

This inference suggests further investigation about the non-significance of predictors other than the age because these non-significant predictors are generally assumed to be very important. Though there is a data limitation but we might try models based on death count data rather than only two categories of death (no death and > 0 deaths). The Poisson and Negative Binomial models are fitted considering this aspect.

4.3. POISSON REGRESSION MODEL:

The Poisson regression model is used treating the response variable number of deaths as count data. The GENMOD procedure of SAS Stat is used to analyze the death count data as a function of Age, Person Year, Gender, Cancer Type and Arsenic Concentration. It is assumed that the response Y (number of deaths) has a Poisson distribution given the predictors. The Age and Person Year were used as 4 categories and quartiles, respectively. Other predictors have the same form as in the logistic regression model. The model was tried with other groups as well such as Age with six categories and Person Year as a continuous variable but, to make the Logistic, Poisson and Negative Binomial models comparable, the same groups were used. The result of the criteria for assessing goodness of fit is presented in Table 4.3.1.

Table 4.3.1: Criteria for assessing goodness of fit of Poisson model

Criterion	DF	Value	Value/df
Deviance	222	284.034	1.279
Pearson Chi-square	222	318.804	1.436
Log Likelihood		172.907	

The mean and variance for Poisson distribution is equal, which implies that the deviance and the Pearson statistic divided by the degrees of freedom should be approximately one. The analysis in Table 4.3.1 shows the value/df for deviance (1.279) and Pearson Chi-square (1.436) is > 1. This indicates the possibility of over-dispersion and hence inadequate fit. This also means that the true variance may be bigger than the mean. The over-dispersion can be tested by fitting a Negative Binomial and comparing its log likelihood with the Poisson model log likelihood. This is the reason why a Negative Binomial model is fitted after a Poisson model. But the values are close to 1 so nothing is ascertained unless tested. The results of the significance test of the Poisson model parameters are presented below in Table 4.3.2.

Table 4.3.2: Results of the significance test of Poisson model

Parameter	Categories	Estimate	$e^{estimate}$	χ^{2a}	P value
Intercept		−2.236		42.48	< 0.0001
Age	20–35	0			
	35–40	1.288	3.625	9.32	0.0023
	40–50	2.425	11.302	51.76	< 0.0001
	>50	3.076	21.672	85.87	< 0.0001
Person Year	Q1	−0.144	0.866	0.60	0.437
	Q2	0.341	1.406	4.04	0.044
	Q3	0.301	1.351	3.72	0.054
	Q4	0			
Gender	Female	0			
	Male	0.269	1.309	10.76	0.001
Cancer Type	Bladder	−0.423	0.655	18.87	< 0.0001
	Liver	−0.494	0.610	24.65	< 0.0001
	Lung	0			
Arsenic Concentration	Low	0.323	1.381	6.40	0.011
	Medium	0.043	1.044	0.14	0.706
	High	0			

a - Wald Chi-square

The parameter estimates given in the table were used to predict the mean number of deaths for different categories of predictors. This is presented in $e^{estimate}$ column. The categories represented as 0 in the estimate column serve as reference categories for other categories of that predictor. Positive estimate parameters suggest that the mean number of deaths is higher for subjects in that category compared to the reference category. In the same way, a negative value suggests that the mean number of deaths is lower for subjects in that category compared to the reference category. The results of the Wald Chi-square tests indicate that there is a statistically significance difference between that category and the reference category. The mean number of deaths can be predicted with the estimates obtained above; for example, for Age Group = 40–50, Person Year = Q3, Gender = Male, Cancer Type = Lung and Arsenic Concentration = Medium,

Predicted death (μ) = exp (−2.236 + 2.425 + 0.301 + 0.269 + 0.423 + 0.043)

All the comparisons of Age group categories (35–40, 40–50 and <50) with reference age group category (20–35) are highly significant. The predicted number of deaths for subjects in Age Group 35–40 is 3.625 or 3.63. This suggests that the subjects in this age group die, on average, 263% more than subjects in the Age Group 20–35. Similarly the subjects in the age group of 40–50 and >50 die, on average, 1030% and 2067% more, respectively, than subjects in the Age Group 20–35. This confirms the results obtained in the logistic model that, as the age increases, the number of deaths increases also. This logically appears correct because the subjects have more exposure to

arsenic if they live longer. Another explanation could be reduction in their immune system as the subjects grew older.

The person year was taken as quartiles in the Poisson regression model. All the first 3 quartiles were compared with 4^{th} quartile to predict the mean number of deaths with reference to the 4^{th} quartile. The comparison of the 1^{st} quartile with the 4^{th} quartile reveals that the subjects in the former die, on average, about 13% less than the subjects in the later category. But this outcome was non-significant ($p = 0.44$). Other comparisons of the 2^{nd} ($p = 0.04$) and 3^{rd} ($p = 0.05$) quartiles with 4^{th} are significant. The subjects in the 2^{nd} and 3^{rd} Person Year quartiles significantly die, on average 41% and 35% more, respectively, than subjects in the 4^{th} quartile. This suggests that as the Person Year increases, the number of deaths decreases, which again confirms the results obtained in bivariate logistic regression model. This inference is based on the significant comparisons of 2^{nd} and 3^{rd} quartiles with 4^{th} quartile. For further investigation of this model, one idea is to have more Person Year categories (Deciles, percentiles or others) and then look at the trend of all the sequential category comparisons.

The comparison of male groups to females is highly significant ($p = 0.001$). This was a big advantage of using the Poisson regression model since the same relationship was non-significant ($p = 0.76$) in the logistic regression. The change of response from two categories to count data shows good gains in the form of more significant relationships. The predicted number of deaths for Males, on average, was 30.9% more than Females. It would be interesting to investigate it further from a medical perspective.

The point of interest could be what biological, physiological or natural changes regarding arsenic intake in males prompts more deaths than in females. It is quite possible that men were exposed more to arsenic compared to women.

The predictor cancer types used in the Poisson regression model were Bladder, Liver and Lung. Here Bladder and Liver cancer were compared with Lung cancer. Both of the comparisons are highly significant. The estimates for Bladder and Liver both are negative, meaning the average deaths for these cancer types are less than for Lung cancer. The subjects affected by Bladder and Liver cancer die, on average, 34.5% (100 - 65.5) and 39% (100 - 61) respectively less compared to Lung cancer. This suggests that the number of Lung cancer deaths is statistically significantly higher compared to Bladder and Liver in the endemic region of Taiwan.

The arsenic concentration was grouped into three levels Low (0–300 ppb), Medium (300–600 ppb) and High (>600 ppb). The Low and Medium concentrations were compared with High. The positive estimate suggests having more deaths in these concentration categories compared to the High concentration. The comparison of Medium vs. High concentration was non-significant ($p = 0.706$) though Low vs. High was significant ($p = 0.011$). The significant Low vs. High comparison suggests that the subjects exposed to Low arsenic concentration die, on average 38.1% more compared to subjects exposed to High concentration. This finding points toward further investigation of the length of exposure, demographic characteristics and other factors in the region.

The overall significance for all the predictors or main effects was tested by Wald statistics for TYPE3 analysis. This is an option provided by SAS Stat package. The results are presented below in Table 4.3.3.

Table 4.3.3: Results of the Wald statistics for Type 3 analysis of Poisson regression model.

Source	DF	χ^2	P value
Age	3	123.44	<.0001
Person Year	3	25.89	<.0001
Gender	1	10.76	0.0010
Cancer Type	2	31.54	<.0001
Arsenic Concentration	2	8.48	0.0144

The Chi-square column entries are likelihood ratio statistics for testing the significance of the predictors added to the model containing all the preceding predictors. The chi-square value of 123.44 for Age represents the difference in log likelihoods between the model with only an intercept term and the model with intercept and Age. The resulting p-value of < 0.0001 indicates that the predictor Age is highly significant. Similarly, the chi-square value of 8.48 for Arsenic Concentration represents the difference in log likelihoods between the Poisson model with the intercept, Age, Person Year, Gender and Cancer Type, and the model with the intercept, Age, Person Year, Gender, Cancer Type and Arsenic Concentration. This effect is also significant as indicated by the p-value 0.0144. The chi-square value suggests the order of importance of the predictors based on their effect on number of deaths as Age ($\chi^2 = 123.44$), Cancer Type ($\chi^2 = 31.54$), Person Year ($\chi^2 = 25.89$), Gender ($\chi^2 = 10.76$), and Arsenic Concentration ($\chi^2 = 8.48$). This finding suggests that Age is the major factor responsible

for severity of arsenic exposure. This also supports the point of keeping the Age variable and dropping Person Year in the final logistic model earlier.

The mathematical form of the Poisson Regression Model can be obtained by putting the value of predicted death μ (explained in Materials and Methods, Section 3.4) in $e^{-\mu}\mu^k/k!$, $k = 0, 1, 2, \ldots$ Depending on the chosen set of variable coefficients, various values of μ result; consequently we only present the general form of the model rather than the numerous possible models that result.

4.4. NEGATIVE BINOMIAL REGRESSION MODEL:

The Deviance (1.279) and Pearson Chi-square (1.436) divided by the degrees of freedom were more than one in the Poisson regression model. This suggests the possibility of over-dispersion, meaning the true variance may be bigger than the mean. The variance and mean are the same for the Poisson distribution and hence, if the above is true then it may be inadequate to fit a Poisson model. The significance of the over-dispersion is tested with the likelihood ratio test based on Poisson and Negative Binomial distribution. This test tests the equality of the mean and the variance imposed by the Poisson distribution against the alternative that the variance exceeds the mean. The test methodology is explained in detail in Materials and Methods section 3.3.2. The test steps are presented below:

H_0: k = 0 (mean = variance: the negative binomial reduces to Poisson)

H_1: k > 0 (mean < variance: the negative binomial does not reduce to Poisson)

Test Steps:

1. Log Likelihood (LL) of Poisson regression model = 172.9073.
2. Log Likelihood (LL) of Negative Binomial (NB) regression model = 175.1062.
3. The likelihood ratio (LR) test statistic is computed as:

   $-2[LL\ (Poisson) - LL\ (NB)] \sim \chi^2_{(1-2\alpha,\ 1)}$

   $= -2[172.9073 - 175.1062] =$ **4.398**

4. $\chi^2_{(.05,\ 1)} = 3.84$, $\chi^2_{(.01,\ 1)} = 6.64$
5. The hypothesis of Poisson distribution is rejected at 5% level of significance and holds at 1% level of significance.

The above finding does not entirely reject (holds at 1%) the validity of doing Poisson regression modeling. However the common and widely accepted norm is a 5% level of significance and the hypothesis of a Poisson distribution does not hold at that level. This also suggests the requirement of more data points or further investigation as this may give a clear direction. The overall conclusion is that both models are fine with this dataset. The criteria for assessing goodness of fit of the Negative Binomial model is presented below in table 4.4.1.

Table 4.4.1: Criteria for assessing goodness of fit of Negative Binomial model

Criterion	Df	Value	Value/df
Deviance	222	243.991	1.099
Pearson Chi-square	222	276.685	1.246
Log Likelihood		175.106	

The assessment based on above statistics is same as presented in the Poisson regression model section 4.3. The values of deviance and Pearson chi-square for Negative Binomial model have substantially reduced from Poisson regression model. This suggests improvement and further validation of the model. The results of the significance test of the Negative Binomial model parameters are presented below in Table 4.4.2.

Table 4.4.2: Results of the significance test of Negative Binomial model

Parameter	Categories	Estimate	$e^{estimate}$	χ^{2a}	P value
Intercept		−2.242		40.64	<.0001
Age	20–35	0			
	35–40	1.294	3.647	9.11	0.0025
	40–50	2.424	11.291	50.13	<.0001
	>50	3.069	21.520	82.32	<.0001
Person Year	Q1	−0.129	0.879	0.38	0.536
	Q2	0.346	1.413	3.22	0.073
	Q3	0.311	1.365	3.15	0.076
	Q4	0			
Gender	Female	0			
	Male	0.277	1.319	8.97	0.003
Cancer Type	Bladder	−0.413	0.662	14.17	0.0002
	Liver	−0.483	0.617	18.58	<.0001
	Lung	0			
Arsenic Concentration	Low	0.333	1.395	5.21	0.023
	Medium	0.039	1.039	0.10	0.754
	High	0			

a - Wald Chi-square

The significance statistic (significant / non-significant) for all the parameter estimates of the predictors except Person Year remains same though probability level

varies from Poisson to Negative Binomial regression model. The 2nd and 3rd quartiles of Person Year turn out non-significant in the Negative Binomial model. The 1st quartile was already non-significant and the 2nd (p = 0.04) and 3rd (p = 0.05) quartiles were borderline significant in the Poisson regression model. The non-significance of Person Year validates dropping it from the logistic model regression earlier. The methodology for predicting the mean number of deaths and the functional form of the Negative Binomial model remains same as explained in the Poisson regression model.

The overall test of significance for all the predictors or main effects of Negative Binomial model were tested by Wald statistics for TYPE3 analysis. This is an option provided by SAS Stat package. The results are presented below in Table 4.4.3.

Table 4.4.3: Results of the Wald statistics for Type 3 analysis of Negative Binomial model.

Source	DF	χ^2	P value
Age	3	112.40	<.0001
Person Year	3	20.25	0.0002
Gender	1	8.97	0.0027
Cancer Type	2	23.37	<.0001
Arsenic Concentration	2	6.83	0.0329

The interpretation remains as explained in the Poisson regression model. The order of importance of predictors remains the same as for the Poisson model. The overall test of significance of all the predictors in Negative Binomial regression model remains same as in the Poisson model though the p value varies.

The mathematical form of the Negative Binomial regression model can be obtained by putting the value of predicted death μ (explained in Materials and Methods, section 3.5) and over-dispersion parameter in $\frac{\Gamma(k+\frac{1}{o})}{k!\Gamma(\frac{1}{o})} = (\frac{o\mu}{1+o\mu})^{k}(\frac{1}{1+o\mu})^{1/o}$. The value of μ varies depending on the coefficients so rather than presenting numerous possible models, we have presented only the general form of the model.

5. SUMMARY AND CONCLUSIONS:

Arsenic is a natural element found in the environment in organic and inorganic form. The inorganic form is much more toxic and is found in ground water, surface water and many foods. This form is responsible for many adverse health effects like cancer and cardiovascular and neurological effects. Arsenic enters into water through industrial waste discharge or particles deposited in dust. The primary source of exposure for most Canadians is food.

Arsenic is a known carcinogen and short/long term exposure could cause abdominal pain; vomiting, diarrhoea, muscular cramping or pain; weakness and skin rash or/and flushing; numbness; burning or tingling sensation or pain in hands and feet; thickening of the skin on the palms and soles, and/or loss of movement and sensory responses; discoloration of the skin; nausea and diarrhoea; decreased production of blood cells; blood vessel damage; and abnormal heart rhythm.

The present study is based on the available mortality data on lung, bladder and liver cancer for the 42 This study suggests significant effect of arsenic in water on mortality. However the study took an epidemiological approach rather than a statistical modelling approach.. In this thesis, we have focussed on further investigation of the data considering a statistical and probabilistic model-based approach. The approach measures the mathematical magnitude and statistical significance of interrelationships between response variable and predictors. The major advantage of this technique is to measure the

likelihood of death by comparing one category of the predictor to another. In this approach one method collapses response into 2 categories viz. no death vs. 1 or more death and another approach considers actual death counts. This study adds knowledge to the existing studies by first exploring the appropriateness of a particular model and than measuring the likelihood of death.

The purpose of the study is to see the effect of various predictors like Gender, Cancer Type (Lung, Liver and Bladder), Arsenic Concentration (High, Medium and low), Age and Person Years on Mortality. Different models were fitted to the secondary data. Three major analytical techniques are used to assess the effect of the independent predictors on dependent response (mortality). These are:

1. Logistic Regression Model,
2. Poisson Regression Model, and
3. Negative Binomial Regression Model

The response variable mortality is considered categorical (no death vs. one or more death) for logistic regression model technique and discrete (actual death count) for the other two models (Poisson and Negative Binomial). For different models all the independent predictors are used as categorical. The variable Person Year was first used as continuous but the LOWESS technique [1] confirms the appropriateness of categories hence it has been used as categorical in our models.

5.1 LOGISTIC REGRESSION:

All the predictors except Age have either been found non significant or dropped from the final logistic regression model. The predictor Person Year has been dropped because of its collinearity with the Age variable. None of the interaction and confounding effects have been found significant. The findings based on Odds Ratio estimates suggest that:

1. Those who are in the age group of 35–40 were 5.4 times more likely to have 1 or more deaths respectively, than those in the age group of 20–35.
2. Those who are in the age group of 40–50 were 27.8 times more likely to have 1 or more deaths respectively, than those in the age group of 20–35.
3. Those who are in the age group of >50 were 114.1 times more likely to have 1 or more deaths respectively, than those in the age group of 20–35.

The mathematical form of the final logistic regression model with mortality as dependent variable and Age as independent variable is:

$$\text{Probability of 1 or more death } [f(z)] = 1/(1+e^{-z})$$

Where $z = -1.094 + 1.681*\text{Age}(35\text{-}40) + 3.326*\text{Age}(40\text{-}50) + 4.737*\text{Age}(>50)$

The inference suggests further investigation about the non-significance of all the predictors except Age. As we are conducting a secondary data analysis we are limited by existing data but we might try models based on actual death counts rather than justtwo

categories of death (no death and > 0 deaths). The Poisson and Negative Binomial models are fitted considering this aspect.

5.2 POISSON REGRESSION MODEL

The significance of the predictors is tested using Wald statistics for this analysis. All the predictors have been found highly significant though there is a possibility of over-dispersion and hence inadequate fit since true variance appears to be greater than the mean. In order to verify the inadequacy of the fit it is necessary to fit the Negative Binomial Model and compare its log likelihood with the Poisson model. This is the reason a Negative Binomial model is fitted to the data. The major inferences deduced from the Poisson Regression model are presented below:

1. The subjects in this age group die, on average, 263% more than subjects in the Age Group 20–35. Similarly the subjects in the age group of 40–50 and >50 die, on average, 1030% and 2067% more, respectively, than subjects in the Age Group 20–35.
2. The subjects in the 2nd and 3rd Person Year quartiles significantly die, on average, 41% and 35% more, respectively, than subjects in the 4th quartile. This suggests that as the Person Year increases, the number of deaths decreases.
3. The predicted number of deaths for Males, on average, was 30.9% more than Females.

4. The subjects affected by Bladder and Liver cancer die, on average, 34.5% and 39% respectively less compared to Lung cancer.
5. The subjects exposed to Low arsenic concentration die, on an average ,38.1% more compared to subjects exposed to High concentration.

The chi-square value suggests the order of importance of the predictors based on their effect on number of deaths as Age (χ^2 = 123.44), Cancer Type (χ^2 = 31.54), Person Year (χ^2 = 25.89), Gender (χ^2 = 10.76), and Arsenic Concentration (χ^2 = 8.48). This finding suggests that Age is the major factor responsible for severity of arsenic exposure. This also justifies dropping of the Person Year not the Age variable from the final logistic model earlier.

5.3 NEGATIVE BINOMIAL REGRESSION MODEL

The over- dispersion of the Poisson distribution model has been statistically tested and found significant at 1% level of significance, not 5%. This suggests the requirement of more data points or further investigation for clear inference. This also validates fitting of both models to the dataset. All the predictors have been found significant and their p values show improvement. The order of importance of predictors based on p values remains same as in the case of Poisson Regression model although the p value varies.

6. REFERENCES

1. Applied Logistic Regression (Second Edition) by David W. Hosmer and Stanley Lemeshow, Published by John Wiley and Sons, Inc, New York.

2. Arsenic in drinking (1999): Subcommittee on arsenic in drinking water committee on toxicology board on environmental studies and toxicology commission on life sciences, National Research Council USA, Published by National Academy Press, Washington, DC.

3. Astolfi, E., S.C. Besuschio, J.C. Garcia-Fernandez, C. Guerra and A. Maccagno. 1982. Hidroarsenicismo Croncio Regional Endemico (in Spanish). Buenos Aires: Cooperativa General Belgrano.

4. Ayerza A. Arsenicismo regional endemico (in spanish). 1917. Bol Acad Nac Med. 1-24.

5. Bates, M.N., A.H. Smith and K.P. Cantor. 1995. Case-control study of bladder cancer and arsenic in drinking water. Am J Epidemiol. 141:523-30.

6. Bates, M.N., A.H. Smith and C. Hopenhayn-Rich. 1992. Arsenic ingestion nd internal cancers: a review. Am J Epidemiol. 135:462-476.

7. Bergoglio, R.M. 1964. Mortalidad por cancer en zones de agues arsenicales de la province de Cordoba, Republica Argentina (in Spanish). Prensa Medica Argentina. 54:994-998.

8. Biagini R.E. 1975. Hidroarsenicismo Croncio en la Republica Argentina (in Spanish). Med Cutan Ibero Latinoam. 423-432.

9. Borgono, P.M., P Vincent, H Venturino et al. 1977. Arsenic in drinking water of the city of Antofagasta: epidemiological and clinical study before and after the installation of the treatment plant. Environ Health Perspect. 19: 103-5.

10. Brown, C.C. and K.C. Chu. 1983. Implications of the multistage theory of carcinogenesis applied to occupational arsenic exposure. J. Natl Cancer Inst., 70, 455.

11. Brown, K.G. K.E. Boyle, C.W. Chen and H.J. Gibb. 1989. A dose-response analysis of skin cancer from inorganic arsenic in drinking water. Risk Analysis, 9(4), 519-528.

12. Buchanan, W.E. 1962. Toxicity of arsenic compound. New York: Elsevier.

13. Cameron, A.C., P.K. Trivedi, Regression analysis of count data, Cambridge University Press, 1998.

14. Cantellano-Alvarado, L., G Viniegra, R.E. Garcia and J.A. Acevedo. 1964. Arsenicismo en la Comarca Lagunera: estudio epidemiologico de arsenicismo en las colonias Miguel Aleman y Eduardo Guerra, de Torreon, Coah (in Spanish). Salud Publica Mex (epoca V). 6:375-385.

15. Cebrian, M.E., A. Albores, M. Aguilar and E. Blakely. 1983. Chronic arsenic poisoning in the north of Mexico. Hum Toxicol. 2:121-133.

16. Chakraborty, A.K., and K.C. Saha. 1987. Arsenical dermotosis from tubewell water in West Bengal. Indian J Med Res. 85:326-34.

17. Chen C.J., C.W. Chen, M.M. Wu and T.-L. Kuo. 1992. Cancer potential in liver, lung, bladder and kidney due to ingested inorganic arsenic in drinking water. Br. J. Cancer, 66, 888-892.

18. Chen C.J., Y.C. Chuang, T.M. Lin, and H.Y. Wu. 1985. Malignant neoplasma among residents of a blackfoot disease-endemic area in Taiwan: High-arsenic artesian well water and cancers. Cancer Res. 45:5895-99.

19. Chen C.J., Y.C. Chuang, S.L. You, T.M. Lin, and H.Y. Wu. 1986. A retrospective study on malignant neoplasm of bladder, lung and liver in blackfoot disease endemic area in Taiwan. Br J Cancer. 53:399-405.

20. Chen C.J., T.L. Kuo, and M.M. Wu. 1988. Arsenic and cancers. The Lancet. 414-15.

21. Chen C.J. and C.J. Wang. 1990. Ecological correlation between arsenic level in well water and age-adjusted mortality from malignant neoplasms. Cancer Res., 50, 5470.

22. Chen C.J., M.M. Wu and S.S. Lee et al. 1988. Atherogenicity and carcinogenicity of high-arsenic artesian well water: multiple risk factors and related malignant neoplasms of blackfoot disease. Arteriosclerosis. 8:452-60.

23. Chen C.W. and Chen C.J. 1991. Integrated quantitative cancer risk assessment of inorganic arsenic. In proceedings of the symposium on Health Risk Assessment on Environmental, Occupational and Lifestyle Hazards, Wen, C.P. (ed.). p.66. Institute of Biomedical Sciences, Academia Sinica: Taipei.

24. Chen K.P. and H.Y. Wu. 1962. Epidemiological studies on blackfoot disease. 2. A study of source of drinking water in relation to the disease. J. Formosam Med. Assoc., 61: 611-618.

25. Chen K.P., H.Y. Wu and T.C. Wu. 1962. Epidemiological studies on blackfoot disease in Taiwan. 3. Physiochemical characteristics of drinking water in endemic blackfoot disease areas. ln: Memoirs, College of Medicine, National Taiwan University, Vol. 8, pp. 115-129. Taipei: National Taiwan University College of Medicine.

26. Ch'i I.C. and R.Q. Blackwell. 1968. A controlled retrospective study of blackfoot disease, an endemic peripheral gangrene disease in Taiwan. Am J Epidemiol. 88:7-24.

27. Chiou, H.Y., Y.M. Hsueh, K.F. Liaw, S.F. Horng, M.H. Chiang, Y.S. Pu. J.S. Lin, C.H. Huang and C.J. Chen. 1995. Incidences of internal cancer and ingested inorganic arsenic: a seven year follow-up study in Taiwan. Cancer Res. 55:1296-1300.

28. Cuzick, J., P. Sasieni and S Evans. 1992. Ingested arsenic, keratoses and bladder cancer. Am J Epidemiol. 136:417-21.

29. Done, A.K. and AJ Peart. 1971. Acute toxicities of arsenical herbicides. Clin Toxicol. 4:343-55.

30. Grobe J.W. 1976. Peripheral circulatory disorders and acrocyanosis in Moselle valley vineyards workers with arsenic poisoning. Berufsdermatosen. 24(3): 78-84.

31. Higgins, I., K. Welch and C. Burchfiel. 1982. Mortality of Anaconda smelter workers in relation to arsenic and other exposures. Department of Epidemiology, University of Michigan: Ann Arbor, MI, USA.

32. Hopenhayn-Rich, C., M.L. Biggs, A. Fuchs, R. Bergoglio, E.E. Tello, H. Nicolli, and A.H. Smith. 1996. Bladder cancer mortality associated with arsenic in drinking water in Argentina. Epidemiology. 7:117-124.

33. International Agency for Research on Cancer. 1980. IARC monograph on the evaluation of carcinogenic risk of chemicals to humans: Some metals and metallic compounds. Lyon, France: World Health Organization. 23:39-141.

34. International Agency for Research on Cancer. 1987. IARC monograph on the evaluation of carcinogenic risks to humans: overall evaluation of carcinogenicity: an updating of IARC monographs volumes 1 o 42. IARC Publ. Suppl., 7, 100.

35. Kadas I., L. Balazs, A Par et al. 1985. Angiosarcoma of the liver following brief arsenic therapy. Zentralbl Allg Pathol. 130:539-43.

36. Lee –Feldstein, A. 1983. Arsenic and respiratory cancer in man: Biomedical and Environmental Perspective, Lederer, W. & Fensterheim, R (eds). Van Nostrand Reinhold: New York.

37. Liang M., L. Zhen-Dong, and Z Ge-you et al. 1995. Clinical analysis and pathogenic changes of skin in endemic chronic arsenicism. Book of posters from the SEGH Second International Conference on Arsenic Exposure and Health Effects, San Diego, CA, June 12-14, 1995. Denver, Colorado: Environmental Science, University of Colorado.

38. Lu F.J. 1975. Physiochemical characteristics of drinking water in blackfoot disease endemic area in chiai and Tainan Hsiens. J. Formosam Med. Assoc., 74: 596-605.

39. Luchtrath, H. 1983. The consequences of chronic arsenic poisoning amomg Moselle wine growers. J Cancer Res Clin Oncol. 105:173-82.

40. Matos, E.L., D.M. Parkin, D.I. Loria, and M. Vilensky. 1990. Geographical patterns of cancer mortality in Argentina. Int J Epidemiol. 19:860-870.

41. Morales, K.H., L. Ryan, K.G. Brown, T.L. Kuo, C.J. Chen, and M.M. Wu. 1999. Model sensitivity in an analysis of arsenic exposure and bladder cancer in southwestern Taiwan. In: Proceedings of the third international conference on arsenic exposure and health effects. 12-15 July 1998, San Diego, California. New York: Elsvier. 201-217.

42. National Research Council. 2001. Arsenic in Drinking Water: 2001 Update.. Subcommittee to update the 1999 Arsenic in Drinking Water Report. National Research council, National Academy of Science. Washington, DC: National Academy Press.

43. Natural Resources Defense Council. 2000. Arsenic and Old Laws.

44. Neubauer, O. 1947. Arsenic cancer: a review. 1962. Br J Cancer. 1:192-251.

45. Roth F. 1957. The sequelae of chronic arsenic poisoning in Moselle vintners. Ger Med Monthly. 2:172-5.

46. Science Daily, Source: Dartmouth Medical School. April 2003. Arsenic in drinking water may be linked to cancer Dartmouth study finds. The study published in the April issue of *International J Cancer.* Authors: Dr. Angeline Andrew, Dr. Margaret Karagas and Dr Joshua Hamilton.

47. Smith, A.H., C. Hopenhayn-Rich, M.N Bates., H.M. Goeden, I. Hertz-Piccioto, H. Duggan, R. Wood, M Kosnett and M.T. Smith. 1992. Cancer risks from arsenic in drinking water. Environ Health Perspectives. 97:259-267.

48. Smith, A.H., M. Goycolea, R. Haque, and M.L. Biggs. 1988. Increase in bladder and lung cancer mortality in a region of northern Chile due to arsenic in drinking water. Am J Epidemiol. 147:660-669.

49. Tello, E.E. 1986. Arsenicismo hidricos: que es el hidroarsenicismo croncio regional endemico Argentino (HACREA) (in Spanish). Arch Argent Dermatol. 36:197-214.

50. Tseng, W.P. 1977. Effects and dose-response relationship of skin cancer and blackfoot disease with arsenic. Environ Health Perspect. 19:109-19.

51. Tseng, W.P., W.Y. Chen, J.L. Sung and J.S. Chen. 1961. A clinical study of blackfoot disease in Taiwan: an endemic peripheral vascular disease. In: Memoirs, College of Medicine, National Taiwan University, Vol. 7, pp. 1-18. Taipei: National Taiwan University College of Medicine.

52. Tseng, W.P., H.M. Chu, S.W. How, J.M. Fong, C.S. Lin and S. Yeh. 1968. Prevalence of skin cancer in an endemic area of chronic arsenicism in Taiwan. J Natl. Cancer Inst. 40(3):453-463.

53. Tsuda, T., A Babazono, E.M. Yamamoto et al. 1995. Ingested arsenic and internal cancer: a historical cohort study followed for 33 years. Am J Epidemiol. 141:198-209.

54. US Environmental Protection Agency. 1984. Health assessment document for inorganic arsenic. EPA 600/8 – 83/021F. Cincinnati, OH: U.S. Environmental Protection Agency.

55. US Environmental Protection Agency. 1984. Health assessment document for inorganic arsenic. Washington DC: Environmental Protection Agency.

56. US Environmental Protection Agency. 1989. Special report on Ingested Inorganic Arsenic: Skin Cancer; Nutritional Essentiality. In: Levine T, Rispin A, Scott C, Marcus W, Chen C, Gibb, H. Risk Assessment Forum. EPA/625:3-87/013. Washington DC: US Environmental Protection Agency.

57. US Environmental Protection Agency, Report to Congress. 2000. EPA studies on sensitive subpopulatios and drinking water contaminants. EPA 815-R-00-015.

58. US Environmental Protection Agency. 2001. National primary drinking water regulations; Arsenic and clarification to compliance and new source contaminants monitoring. Final Rule. Federal Register 66 (14), 6976-7066.

59. Vahter, M. 1994. Review: Species difference in the metabolism of arsenic compounds. Applied Organometallic Chemistry. 8(3), 175-182.

60. Vallee B.L., DD Ulmer and WEC Wacker. 1960. Arsenic toxicology and biochemistry. Arch Ind Health. 21:132-51.

61. Virarghavan, T., K.S. Subramaniam and T.V. Swaminathan. 1994. Drinking water without arsenic: A review of treatment technologies. Environmental Systems Reviews, 37.

62. World Health Organisation. 1981. International program on chemical safety, environmental health criteria 18: Arsenic. Geneva: WHO.

63. Wu CM. 1961. Geologic studies on ground water in blackfoot disease endemic area. In: Report on Blackfoor Disease Research, Vol. 5, pp. 1-24. Taichung: Taiwan Provincial Department of Health.

64. Wu, H.Y., K.P. Chen, W.P. Tseng and C.L. Hsu. 1961. Epidemiological studies on blackfoot disease. 1. Prevalence and incidence of the disease by age, sex, year, occupation, and geographic distribution. In: Memoirs, College of Medicine, National Taiwan University, Vol. 7, pp. 33-50. Taipei: National Taiwan University College of Medicine.

65. Wu, M.M., T.L. Kuo, Y.H. Hwang, and C.J. Chen. 1989. Dose-response relation between arsenic concentration in well water and mortality from cancers and vascular disease. Am. J. Epidemiology. 130(6):1123-1132.

66. Zaldivar R. 1974. Arsenic contamination of drinking water and food-stuff causing endemic chronic poisoning. Beitr Pathol Bd. 151:384-400.

APPENDIX 1

Table A.1 Internal Cancer Incidence by Age and by Arsenic Concentration Group [65][1]

Age Group (yr)	Arsenic Concentration, ppb											
	0–300				300–600				>600			
	PY[2]	lng[3]	bl[4]	liv[5]	PY[2]	lng[3]	bl[4]	liv[5]	PY[2]	lng[3]	bl[4]	liv[5]
Male												
20–25	35,421	0	0	0	17,754	0	0	0	10,477	0	0	0
25–30	21,439	0	0	0	9,802	0	0	0	6,132	1	0	0
30–35	13,493	0	0	2	6,356	0	0	2	4,507	0	1	2
35–40	12,432	0	0	4	6,000	1	0	2	3,591	0	0	2
40–45	13,550	2	1	3	6,765	2	2	3	3,852	0	1	3
45–50	13,395	4	0	5	6,423	6	3	5	3,823	3	2	5
50–55	11,293	7	2	6	5,507	5	4	3	3,115	5	3	3
55–60	8,934	7	3	10	4,276	11	4	5	2,482	10	6	5
60–65	7,020	5	3	7	3,431	10	4	4	1,828	4	3	4
65–70	5,229	7	5	9	2,533	4	3	3	1,148	4	3	3
70–75	3,676	15	2	4	1,695	8	5	1	748	3	2	1
75–80	2,005	7	5	3	883	5	4	0	317	3	5	0
80–85	1,190	5	5	1	643	1	3	1	159	0	1	1
Female												
20–25	27,908	1	0	0	13,131	0	0	0	8,442	0	0	0
25–30	15,107	0	0	0	6,799	0	0	0	4,546	0	0	0
30–35	11,600	1	0	0	5,145	0	0	0	3,800	0	0	0
35–40	11,932	0	1	0	5,759	0	0	1	3,612	1	0	1
40–45	13,373	5	0	2	6,774	3	0	1	4,014	1	0	1
45–50	13,109	3	2	2	6,665	3	0	2	4,114	4	0	2
50–55	11,368	7	2	1	5,708	4	3	1	3,512	5	4	1
55–60	9,241	3	4	3	4,616	6	3	1	2,571	11	3	1
60–65	7,753	8	9	5	3,732	4	3	2	1,800	9	10	2
65–70	5,998	10	3	3	2,825	6	10	3	1,201	3	4	3
70–75	4,198	3	5	4	1,907	6	5	1	668	2	3	1
75–80	2,323	5	4	2	1,154	2	2	0	352	1	3	0
80–85	1,860	2	2	5	787	2	4	0	23	1	1	0

[1]Data from Wu et al. 1989; Chen et al. 1992.
[2]Person-years at risk.
[3]Number of death from lung cancer.
[4]Number of death from bladder cancer.
[5]Number of death from liver cancer.