INVESTIGATION OF CANCER DEATH RISK IN THE COMORBIDITY CASE
ИЗУЧЕНИЕ РИСКА СМЕРТИ ОТ РАКА ПРИ НАЛИЧИИ СОПУТСТВУЮЩИХ ЗАБОЛЕВАНИЙ

Tsurko V.V., Michalski A.I.

Abstract

In this paper we investigate dependences between associated diseases that a person have at the end of his live and the cause of death. We analyze public data about cause-specific mortality in conjunction with the problem of generalization risk estimation on empirical data. The use of the theory of Vapnik-Chervonenkis provides informative results about differences between distributions of associated diseases in group of people who died of cancer and group of people who died of another disease, uncovers a relationship between some groups of associated diseases and risk of death of cancer.

Keywords: cancer mortality, distributions discrepancy, selection of associated diseases, the Vapnik-Chervonenkis dimension

Introduction

Investigation of relationships between health and mortality is relevant because of longevity increase, which is observed in developed countries starting from the second part of the XX century. In order to release economic and social pressure due to ageing of population it is important to have the solution to following problems: obtaining reliable estimates for expected age structure of the population, gaining knowledge of factors responsible for "healthy aging" and understanding the impact of different diseases in cause-specific mortality. The last problem is known as mortality-comorbidity problem. This problem is especially important for old age groups in which the mortality is at the high level and several chronic diseases are presented.

Nowadays there is a great volume of statistical data for mortality and morbidity of the aged people. These data allow us to investigate the factors responsible for maintaining health in aging population, to evaluate the influence of heredity, environment and lifestyle. Recent publications explain observed increase of human life expectancy by the reduction of mortality at middle age [2, 3, 4]. There is a hypothesis that people who are down in health have a high margin of "active longevity", because enduring high morbidity risk in young and middle ages gives an advantage in survival in old age. Effective adaptation of people with chronic diseases may serve as a biological basis for this phenomenon. If this hypothesis is correct, we need to focus preventive measures to ensure the "healthy aging" in the age groups of young and middle ages. Relationship between the cause specific mortality and chronic diseases can be an indirect confirmation of the relationship between increased morbidity and reduced mortality. In the research the Multiple Cause-of-Death Public-Use Data for 2007 by the National Center for Health Statistics USA [5] are investigated. Distribution of associated diseases presented by the ICD10 codes among people who died of cancer (C00-C97) is compared with the same distribution among people who died of another disease. In order to select more "important" diseases associated with cancer mortality we solve a problem of contrasting the distributions. By the problem of contrasting of distributions we mean the selection of associated diseases for which we have the most distinguishable distributions of these diseases among people died of cancer against people died of other diseases.

We used symmetrized Kullback-Leibler divergence as a difference measure between the two distributions. For a set of associated diseases the symmetrized Kullback-Leibler divergence was estimated from the data as a half sum of mixed entropies corrected by a penalty term. This term takes into account both the amount of empirical data and the number of considered associated diseases. In this report the technique for construction of such penalty term based on the Vapnik-Chervonenkis dimension was used.

The results show that in a group of women at the age of 45-75 years, which died of cancer, not more than 9 of 126 classes of associated diseases are linked with cancer mortality. Considering partition into groups of diseases we conclude that the class of chronic pulmonary and respiratory diseases is one of the most important in cancer and non cancer comorbidity. This research validates the hypothesis that presence of asthma could decrease development and mortality of cancer [3].

Definitions

In this section we introduce some definitions and notation that will be used throughout the paper.

We consider the problem of approximation a distance between two distributions $p_1(x)$ and $p_2(x)$ on empirical data, where $p_1(x)$ is a distribution of associated diseases among the group of people who died of cancer (let's name this group as a cancer group), $p_2(x)$ - a distribution of associated diseases among the group of people who died of another disease (non cancer group). Associated diseases are grouped into classes according their ICD10 classification, x is a group of associated diseases.
We used the symmetrized Kullback-Leibler divergence as a distance between empirical estimates $\hat{p}_1(x)$ and $\hat{p}_2(x)$ of the two distributions:

$$D = -\frac{1}{2}\left(\sum_x p_2(x)\ln\frac{\hat{p}_1(x)}{p_2(x)} + \sum_x p_1(x)\ln\frac{\hat{p}_2(x)}{p_1(x)}\right)$$

The value D is minimal if $\hat{p}_1(x) \equiv p_1(x)$ and $\hat{p}_2(x) \equiv p_2(x)$. Because distributions $p_1(x)$ and $p_2(x)$ are unknown, we will estimate the distance D on empirical data.
For each block of associated diseases values n_i and m_i are calculated, where n_i is the number of people in the cancer group which had a disease from the *i*th block, m_i – the number of people in the non cancer group which had a disease from the *i*th block. The cancer and non cancer groups have histograms of associated diseases: $g_1 = (n_{(1)}, n_{(2)}, \dots, n_{(k)})$ and $g_2 = (m_{(1)}, m_{(2)}, \dots, m_{(k)})$, where k is the number of blocks and these blocks are sorted in descending order of the absolute difference between values $n_i/\sum_{i=1}^k n_i$ and $m_i/\sum_{i=1}^k m_i$.
To find a set of associated diseases which are the most important for the difference between cancer and non cancer death we consider different sets of blocks of associated diseases. Let α be a variable which labels what set of blocks we use now, Σ is the set of all possible sets α. In an experimental part we create the following sequence of sets α: $\alpha_{(1)}$ - the first block of associated disease, $\alpha_{(2)}$ - the first and the second blocks, … , $\alpha_{(k)}$ - all blocks, where the order of blocks is the same as in histograms above. Distributions of associated diseases now depend on the variable α, and the symmetrized Kullback-Leibler divergence takes form

$$D(\alpha) = -\frac{1}{2}\left(\sum_x p_2(x,\alpha)\ln\frac{\hat{p}_1(x,\alpha)}{p_2(x,\alpha)} + \sum_x p_1(x,\alpha)\ln\frac{\hat{p}_2(x,\alpha)}{p_1(x,\alpha)}\right)$$

In the rest of the article we consider a functional of an average risk as a characterizing criterion of the distance $D(\alpha)$:

$$M(\alpha) = -\frac{1}{2}\left(\sum_x p_2(x,\alpha)\ln\hat{p}_1(x,\alpha) + \sum_x p_1(x,\alpha)\ln\hat{p}_2(x,\alpha)\right) \quad (1)$$

The distributions $p_1(x,\alpha)$ and $p_2(x,\alpha)$ are unknown and are approximated by frequencies. We can use a trivial approximation by frequencies $\nu_1(x,\alpha)$ and $\nu_2(x,\alpha)$ which are equal to a portion of people who had an associated disease x and died of cancer or of another disease respectively. If x is an ith block of associated diseases and α consists of k blocks then frequencies are defined as:

$$\nu_1(x,\alpha) = n_i/\sum_{i=1}^k n_i, \nu_2(x,\alpha) = m_i/\sum_{i=1}^k m_i$$

To avoid a zero value under logarithm in (1) we use the empirical estimates $\hat{p}_1(x,\alpha)$ and $\hat{p}_2(x,\alpha)$ of distributions $p_1(x,\alpha)$ and $p_2(x,\alpha)$ in form

$$\hat{p}_1(x,\alpha) = \frac{n_i+1}{\sum_{i=1}^k n_i + k}, \hat{p}_2(x,\alpha) = \frac{m_i+1}{\sum_{i=1}^k m_i + k} \quad (2)$$

These expressions are Bayes estimates of probabilities if a priori distributions of probabilities on the k-fold simplex given by $\Delta^k = \{p_1, \dots, p_k : \sum_{i=1}^k p_i = 1, p_i \geq 0, i = 1, \dots, k\}$ are uniform.
By substitution of $\nu_1(x,\alpha)$ and $\nu_2(x,\alpha)$ instead of $p_1(x,\alpha)$ and $p_2(x,\alpha)$ in (1) we obtain so call empirical risk

$$M_e(\alpha) = -\frac{1}{2}\left(\sum_x \nu_2(x,\alpha)\ln\hat{p}_1(x,\alpha) + \sum_x \nu_1(x,\alpha)\ln\hat{p}_2(x,\alpha)\right) =$$

$$= -\frac{1}{2}\left(\frac{1}{\sum_{j=1}^k m_j}\sum_{i=1}^k m_i \ln\frac{n_i+1}{\sum_{j=1}^k n_j + k} + \frac{1}{\sum_{j=1}^k n_j}\sum_{i=1}^k n_i \ln\frac{m_i+1}{\sum_{j=1}^k m_j + k}\right) \quad (3)$$

The deviation between the average risk and the empirical risk can be estimated in form of inequality

$$M(\alpha) > M_e(\alpha) - d(\alpha,\eta),$$

which is valid with probability η.

By maximizing the right part of the inequality we determine the set of classes of associated diseases for which the distribution of associated diseases in the cancer group maximally differs from the distribution of associated diseases in the non cancer group. The form of the penalty term $d(\alpha,\eta)$ and some empirical results are discussed in the rest of the article.

Vapnik-Chervonenkis evaluation

In this section we consider functionals of the average and empirical risks and discuss an applicability of the Vapnik-Chervonenkis evaluation as a bound of the difference between these risks.

We consider the functional of the average risk $M(\alpha)$ in form (1) and the functional of the empirical risk $M_e(\alpha)$ in form (3).

Let $x_{1i}^{\alpha}, i = 1,\dots,L_1^{\alpha}$ denote a block of associated diseases which *i*th person from the cancer group had, where L_1^{α} is the number of people who belonged to the cancer group and had an associated disease from a set α. In the same way, let $x_{2i}^{\alpha}, i = 1,\dots,L_2^{\alpha}$ denote a block of associated diseases which *i*th person from the non cancer group had, where L_2^{α} is the number of people who belonged to the non cancer group and had an associated disease from a set α.Then we can obtain the following expression for the empirical risk (3):

$$M_e(\alpha) = -\frac{1}{2}\left(\frac{1}{L_2^{\alpha}}\sum_{i=1}^{L_2^{\alpha}} \ln\hat{p}_1(x_{2i}^{\alpha},\alpha) + \frac{1}{L_1^{\alpha}}\sum_{i=1}^{L_1^{\alpha}} \ln\hat{p}_2(x_{1i}^{\alpha},\alpha)\right) \quad (4)$$

We want to use the Vapnik-Chervonenkis result from [1] about the uniform convergence of means to expectations in class of bounded functions. The result is as follows: assume $F(x,\alpha)$ is a measurable function for all $\alpha \in \Sigma$ with respect to $P(x)$ in probability space X, $M(\alpha)$ is an expectation of this function for all α

$$M(\alpha) = EF(x,\alpha) = \int F(x,\alpha)dP(x)$$

Then assume an independent sample with the distribution $P(x)$: $X^l = x_1,\dots,x_l$. For all α we calculate an average value of $F(x,\alpha)$ for X^l

$$M_e(\alpha) = \frac{1}{l}\sum_{i=1}^{l} F(x_i,\alpha)$$

If the function $F(x,\alpha)$ is bounded: $0 \le F(x,\alpha) \le a$ then the inequality holds

$$P\left\{\sup_{\alpha\in\Sigma}|M(\alpha) - M_e(\alpha)| > a\varepsilon\right\} \le 6M^S(2l)\exp\left[-\frac{1}{4}\varepsilon^2(l-1)\right], \quad (5)$$

where $M^S(2l)$ is the growth function of a system of events $A = \{x: F(x,\alpha) \ge c, c > 0, \alpha \in \Sigma\}$. The proof of (5) and the definition of the function $M^S(2l)$ are given in [1]. From (5) it follows that with probability not less than $1-\eta$ for any $\alpha \in \Sigma$ the following inequality holds

$$|M(\alpha) - M_e(\alpha)| \le 2a\sqrt{\frac{\ln 6M^s(2l) - \ln\eta}{l-1}} \quad (6)$$

To use this inequality let's prove that the function $F(x,\alpha) = -\ln\hat{p}_i(x,\alpha)$ is bounded. Hence, we should prove two inequalities:

$a_1 \le \ln\hat{p}_i(x,\alpha) \le a_2, \qquad i = 1,2$

Indeed, we don't consider associated diseases which neither people from the cancer group nor from the non cancer group had. Then we have $\hat{p}_i(x,\alpha) > c > 0$ and $-\ln\left(\frac{\hat{p}_i(x,\alpha)}{ce}\right) < 1$. By the definition (2): $\hat{p}_i(x,\alpha) < 1$, and therefore $-\ln\left(\frac{\hat{p}_i(x,\alpha)}{ce}\right) > \ln(ce)$.

So we've proved the inequality:

$$\ln(ce) < -\ln\left(\frac{\hat{p}_i(x,\alpha)}{ce}\right) < 1$$

Functions $\ln\left(\frac{\hat{p}_i(x,\alpha)}{ce}\right)$, $i = 1,2$ are bounded and the equality holds:

$$-\frac{1}{2}\left(\frac{1}{L_2^\alpha}\sum_{i=1}^{L_2^u}\ln\left(\frac{\hat{p}_1(x_{2i}^\alpha,\alpha)}{ce}\right)+\frac{1}{L_1^\alpha}\sum_{i=1}^{L_1^u}\ln\left(\frac{\hat{p}_2(x_{1i}^\alpha,\alpha)}{ce}\right)\right)$$

$$= \ln(ce) - \frac{1}{2}\left(\frac{1}{L_2^\alpha}\sum_{i=1}^{L_2^\alpha}\ln\hat{p}_1(x_{2i}^\alpha,\alpha)+\frac{1}{L_1^\alpha}\sum_{i=1}^{L_1^\alpha}\ln\hat{p}_2(x_{1i}^\alpha,\alpha)\right)$$

Now we can use the estimation (6) for the empirical risk. With the estimation $M^S(2l) \leq (2l)^k$ this leads to the inequality which holds with probability not less than $1-\eta$ for all sets of associated diseases composed not more than k classes

$$M(\alpha) > M_e(\alpha) - 2\sqrt{\frac{2^{k-1}\left(\ln\frac{L_1^\alpha + L_2^\alpha}{2^{k-2}}+1\right)-\ln\frac{\eta}{5}}{L_1^\alpha + L_2^\alpha - 1}} \qquad (7)$$

Experimental results

In this section we present the analysis of the data about human comorbidity and mortality. We are interested in differences between two groups of people: people who died of cancer and people who died of another disease. Usually a person in addition to underlying disease (the case of death) has a list of associated diseases. Therefore there are certain distributions of associated diseases in these two groups of people.

For the analysis the Multiply Cause-of-Death Public-Use data for 2007 are used. These data contain the information about people who died in 2007 year. About each person we have: age, date of death, a disease which was a cause of death, a list of associated diseases. We made our differences analysis on the data of women morbidity.

We consider the age range in which the cancer mortality is the most common. Figure 1 shows a percentage of deaths of cancer depending on age. The horizontal axis corresponds to an age of death; the vertical axis corresponds to a percentage of cancer deaths among women died at the certain age. From the figure 1 we see that the majority of cancer deaths (more than 30%) are in the age interval between 45 and 75 ages. In the rest of the article we consider the group of women who died at the age interval 45-75 years.

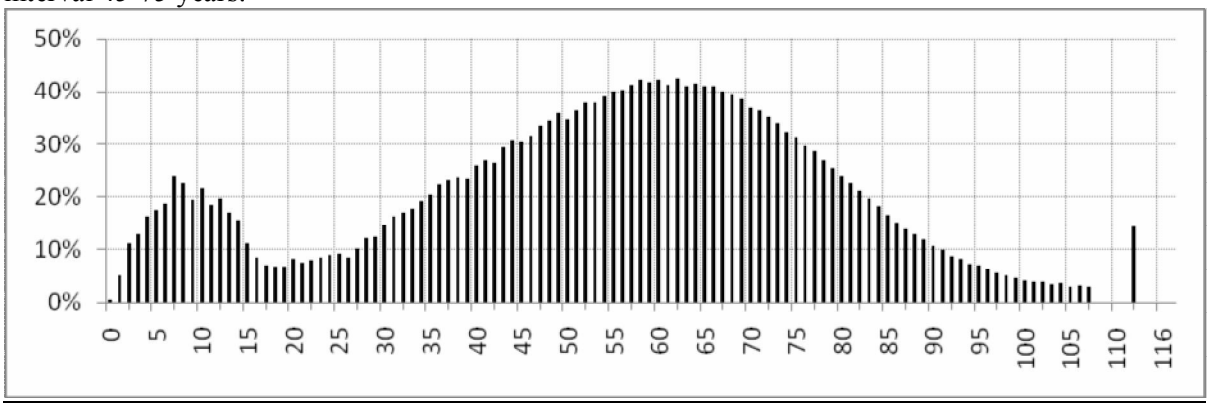

Figure 1. Histogram of proportion of women died of cancer

At the first step of our analysis we consider 24 classes of associated diseases; these classes correspond to the first letter in the ICD10 code.
A number of women from the cancer or non cancer group who had the associated disease from fixed class is calculated (according to our definitions n_i and m_i respectively). Classes of associated diseases are sorted in descending order of the absolute difference between values $n_i / \sum_{i=1}^{k} n_i$ and $m_i / \sum_{i=1}^{k} m_i$.
To evaluate the empirical and average risks for the experimental data and to find a set of associated diseases which are the most important for difference between cancer and non cancer death we consider different sets α of classes of associated diseases. We create follow sequence of sets α: $\alpha_{(1)} = \{\text{J}\}$ is diseases of respiratory system, $\alpha_{(2)} = \{\text{J,I}\}$ - diseases of respiratory system and circulatory system, $\alpha_{(k)}$ – all considered classes, where the order of classes is the same as defined above.
Using the mortality data we calculate values of the empirical risk functional of the form (4) for each set α. According to inequality (7) we evaluate the lower bound of the average risk. In figure 2 the empirical risk $M_e(\alpha)$ and the lower bound of the average risk $M(\alpha)$ are plotted. The lower bound of the average risk reaches its maximum on the set $\alpha = \{\text{J,I, E, D, R, N}\}$, so we've determined the set of classes of associated diseases for which the distribution of associated diseases in the cancer group maximally differs from the distribution of associated diseases in the non cancer group. This result shows the relationship with cancer mortality and the importance of such classes of associated diseases as: Diseases of the respiratory system (J), Diseases of the circulatory system (I), Endocrine, nutritional and metabolic diseases (E), Neoplasms and diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D), Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (R), Diseases of the genitourinary system (N).

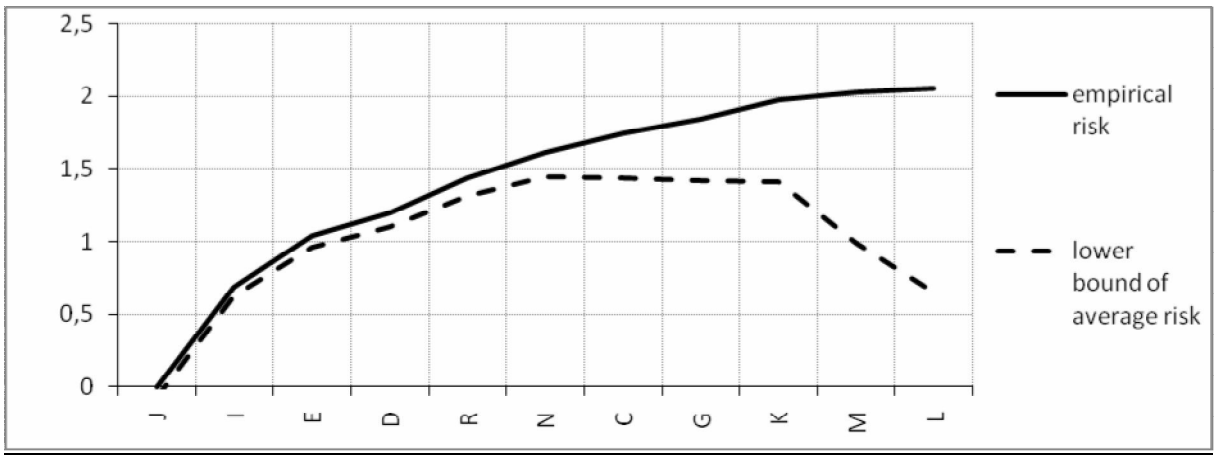

Figure 2. Empirical and average risks for classes of associated diseases

The classes of associated diseases combined by the first letter of the ICD10 code are too large and heterogeneous. At the second step of our analysis we consider more detailed composition of diseases classes: now we use blocks of diseases defined by letter and two digits of the ICD10 code. These blocks are defined in standard classification.
For these blocks of associated diseases we perform the same comparison of the empirical and average risks. From figure 3 one can conclude that the cancer group maximally differs from the non cancer group on nine blocks of associated diseases. We emphasize such diseases as: Ischemic heart diseases (I20-I25), Hypertensive diseases (I10-I15), Other diseases of the respiratory system (J95-J99), Chronic lower respiratory diseases (J40-J47), Circulatory and respiratory systems (R00-R09), Renal failure (N17-N19), Diabetes mellitus (E10-E14), Obesity and other hyperalimentation (E65-E68), Influenza and Pneumonia (J09-J18). Some of these diseases may play protective role against cancer death, some can be artifacts. A part of the found relationships between cancer and associated diseases are well-known, some of these relationships are being discussed in professional area [6].

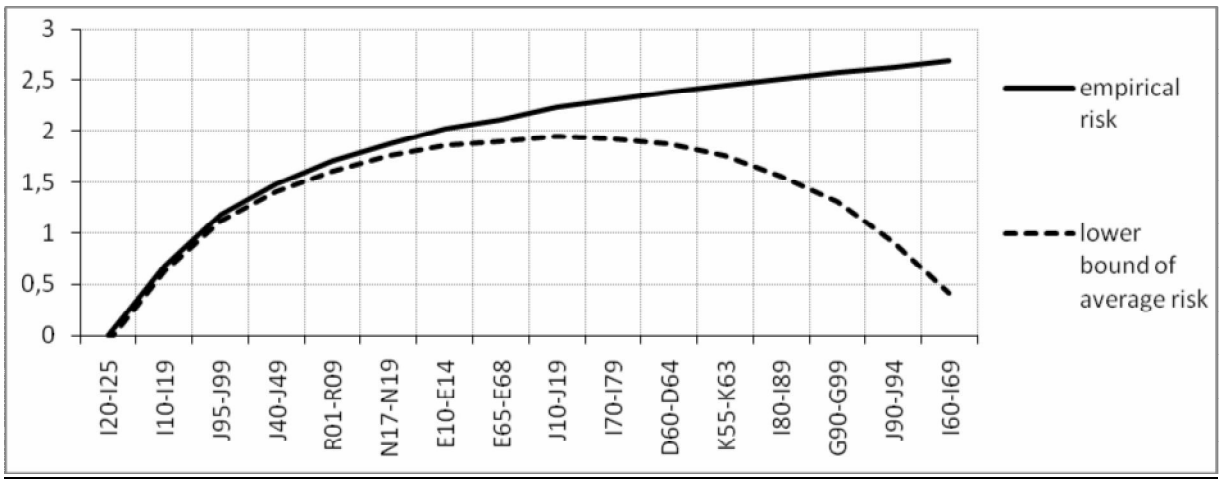

Figure 3. Empirical and average risks for blocks of associated diseases

Conclusion

This paper is devoted to the problem of investigation of links between risk of cancer death and associated morbidity. It is mathematically formalized as the problem of contrasting the distributions of associated diseases among people died of cancer and among people died of another disease. To solve this problem we evaluate the average risk on the empirical data using the Vapnik-Chervonenkis inequalities. We perform two partitions of groups of diseases and obtain the better, more interpretable results on the second partition (letter and two digits of the ICD10 code). It turns out that nine blocks of associated diseases have reliably different distributions in the cancer and non cancer groups. This allows us to discuss the role of some chronic diseases and conditions in cancer mortality.

The aim of the future investigations is consideration of the more "tiny" classes of associated diseases that reduce cancer mortality. For such classes one should use more precise estimation for the average risk than estimation based on the Vapnik-Chervonenkis approach.

References

[1] V. Vapnik, "Statitical Learning Theory", *Wiley Interscience, 1998*

[2] M.V. Blagosklonny, "Why human lifespan is rapidly increasing: solving "longevity riddle" with "revealed-slow-aging" hypothesis", *Ag- ing, vol. 4, p.177-182, 2010*

[3] A.I. Michalski, S.V. Ukraintseva, K.G. Arbeev, A.I. Yashin, "Investigation of old age mortality structure in the presence of comorbidity. In European Conference On Chronic Disease Prevention", *Helsinki, Finland, p.60, 2005*

[4] A.I. Yashin, et al, "Have the oldest old adults ever been frail in the past? A hypothesis that explains modern trends in survival", *J. Gerontol. Biol. Sci., vol.56, p.B432-B442, 2001*

[5] Mortality Data, Multiple Cause-of-Death Public-Use Data Files, *http://www.cdc.gov/nchs/products/elec_prods/subject/mortmcd.htm*

[6] Elmer W. Fisherman M.D., "Does the allergic diathesis influence malignancy?", *Journal of Allergy Volume 31, Issue 1, January-February 1960, Pages 74-78*