Written: 14. Sept 2003
Updated: September 30, 2003
StatisticsRough.doc

Preliminary Statistical Evaluation of the 5 Cancer Cases

Michael Peleg

This analysis is preliminary and needs to be refined and verified.

Facts:

1. The site is distinct by frequent and long term exposure to diverse forms of non-ionizing electromagnetic radiation. The exposure was probably within legal (ICNIRP) limits most of the time.
2. 5 workers working at the site were diagnosed with cancer. The ages at diagnosis were: 34, 36, 39, 40, and 48.
3. Periods of time (years) each of the above workers spent at the site before diagnosis were approximately: 11 nearby, part of them on the exact site, 8, probably more than 3, 9 and 17 respectively.
4. Number of workers, denoted by N, who worked at the site more than 2 years over the relevant period of about 12 years is estimated between 20 to 50, best estimate 30, almost sure not more than 40.

The statistics used

The statistics used has to be rechecked, it was gathered from the internet without the required medical expertise:

Probability of the general population to be diagnosed with cancer from birth to age 40: P1=0.016 (**1.6%**)
Probability of the general population to be diagnosed with first cancer from age 41 to 60: P2=0.085 (**8.5%**)

Probability of the general population not to be diagnosed with cancer from birth to age 60: P3=1-P1-P2

The questions dealt with:

1. What is the statistical p-value? That is:
If a group of N (20 to 50) people is chosen at random from the general population what is the probability, denoted as Pt, that at least 4 of them will be diagnosed with cancer up to age of 40 and at least one of them up to age of 60?

2. What was the risk ratio in the group of workers of being diagnosed with cancer up to the age of 40 (the number of cases relative to that expected in normal population) and what is the corresponding 95% confidence interval?

Analysis

See the section "statistical reference" below.

The p-value, that is, the probability Pt is given by

$$Pt = \sum_{N1=4}^{N} \sum_{N2=1}^{N-N1} P(N; N1, N2, N - N1 - N2; P1, P2, P3)$$

where P is defined in the section "statistical reference" below.

The risk ratio among the group of workers relative to the general population of being diagnosed with cancer up to the age of 40 is denoted by RR. Its 95% confidence interval CI was computed along the lines suggested in http://www.medepi.org/epitools/ , further verification would be welcome.

Results:

N	25	**30**	35	40	50
Pt (p-value)	0.00054 1:1800	**0.0012 1:860**	0.0022 1:461	0.0036 1:275	0.0083 1:120
RR (Risk Ratio)	10	**8.3**	7.1	6.25	5
CI 95%	2.8 – 22.5	**2.3 - 19**	2 – 16.7	1.7 – 14.8	1.4 - 12

That is, for population of N=30, the probability of this occurring at random is 1:860, the risk ratio is 8.3 and its 95% confidence interval is 2.3 to 19.

Further data desired

This analysis should be refined by:

1. Checking expertly the data obtained from internet and obtaining statistics with better resolution than 20 years.
2. Obtaining and incorporating data about the ages of the exposed population and maybe about the specific cancers.

Discussion

The p-value and the risk ratio are certainly statistically significant and indicate increased risk of cancer for young workers exposed to electromagnetic radiation.

There is a possibility of selection bias since this analysis was done on the affected group of workers. This may be partly offset by the following considerations:

1. The working site is very distinct by its radiation, maybe one of a kind, the chance to choose a site of this character at random is very small.

2. The study was initiated after only 3 cases were identified. During the short course of data gathering additional 2 cases were found.
3. The site was not selected automatically by the cancer cases. Two more events had to happen to bring it to the scientific literature. First, somebody had to become aware of the abnormality. This happened years after the last case, due to an unrelated event in the organization. The occurrence of such group of cancer cases is not very obvious due to small number of cases dispersed over many years, some occurring after the affected people moved to other diverse locations. Second, passing this information to scientific literature involves complex processes with uncertain outcome and cannot be taken for granted either. Thus the selection was not solely by the occurrence of the cases and very possibly there are other unreported occurrences elsewhere.

Other possible causes of the cancer cases were not investigated, however no abnormal cancer cases are known among people who worked nearby for many years, including other parts of the same building with lower radiation exposure.

A more expert analysis considering the specific types of cancer, their statistics and rate of occurrence is expected to yield even stronger results.

This study raises serious suspicions about radiation effects and may contribute to more definite conclusions when examined together with similar results occurring elsewhere.

Statistical reference:

Papoulis:" Probability, random variables and stochastic processes", eq. (3-62) on generalized Bernoulli trials:

$$P(N;N1...Nk;P1...Pk) = N!\prod_{i=1}^{k}\frac{Pi^{Ni}}{Ni!}$$

Where N is the number of experiments (population size at our case), k is the number of possible, mutually exclusive, outcomes (age groups at diagnosis in our case), Ni are the numbers of experiments with the different outcomes (numbers of cancer cases at each age group in our case), Pi are the probabilities of those outcomes and ! denotes factorial.
This equation is applied conservatively due to the uncertainty about the ages of the workers. Each of the N workers is associated with a statistical experiment with three possible outcomes: Diagnosed with cancer at age up to 40; diagnosed with cancer at age in the range 40 to 60; not diagnosed with cancer until age of 60. Since the statistics used where the probabilities over whole lifetime and the observation period was 10 to 20 years, the exact p-value is even lower (more significant) then the one calculated here.

Matlab files radstat.m, present.m, confidence.m and genBern.m contain the calculations above. Present.m is the main routine.