Chinese Journal of Cancer Research 9(1):68-75,1997.

A Microcomputer Program in Cancer Registry

STATISTICAL TESTS FOR THE COMPARISON OF THE INCIDENCE OR MORTALITY RATES IN CANCER REGISTRY AND DESCRIPTIVE EPIDEMIOLOGY ——A MICROCOMPUTER PROGRAM IN BASIC

Xiang Yongbing 项永兵　Jin Fan 金凡　Gao Yutang 高玉堂

Department of Epidemiology, Shanghai Cancer Institute, Xie Tu Road 220032

This paper describes the statistical methods of the comparison of the incidence or mortality rates in cancer registry and descriptive epidemiology, and the features of microcomputer program (CANTEST) which was designed to perform the methods. The program was written in IBM BASIC language. Using the program CANTEST we presented here the user can do several statistical tests or estimations as follow: 1. the comparison of the adjusted rates which were calculated by directly or indirectly standardized methods, 2. the calculation of the slope of regression line for testing the linear trends of the adjusted rates, 3. the estimation of the 95% or 99% confidence intervals of the directly adjusted rates, of the cumulative rates (0–64 and 0–74), and of the cumulative risk. Several examples are presented for testing the performances of the program.

Key words: Cancer registry, BASIC, Microcomputer program, Incidence, Mortality, Descriptive epidemiology, Statistical tests.

Cancer registry is a very important work in the fields of epidemiology research, health care planning, and cancer control.[1,2] For data analysis of cancer registration including incidence, mortality and survival, the computer or computer program is a very useful tool for storing, manipulation or processing and statistical analysis. For statistical analysis of cancer incidence or mortality, we often tabulate the number of cancer cases by sex, age groups and primary sites for one or several years in a specified areas, and calculate the rates of age-specific and age-standardized (e.g. World), or other rates. The secular trends of incidence or mortality rates with time were often reported when the data sets collected by registry covered many years.[3] For testing the time trends of incidence or mortality rates, the statistical models of Age-Period-Cohort Models[3-5] are frequently used by statistician. The tests of hypothesis on adjusted rates are one of most important parts in statistical analysis of cancer registration data. But in some computer software or package, there are not the functions for doing the statistical tests of the adjusted rates, such as the CANREG[1,2] developed by IARC. Recently, Immonen-Raiha P, et al.[6] developed a useful SAS macro for calculating the age-standardized incidence rates. But it can not be used to do statistical tests or comparison of age-standardized rates.

In the present paper we describe a microcomputer program CANTEST which can be used to estimate the confidence intervals and test the linear

Accepted December 8, 1996

trends of the adjusted rates, test the hypothesis on comparison of two adjusted rates, and calculate the cumulative risk. The statistical methods in our program are reviewed in the next section, In section of "Program Description" the features of program and requirement we will described. Some examples are given to illustrate the performances of the program in section of "Applications".

STATISTICAL METHODS

There are often differences between the age structures of populations to be compared, or when describing the secular trends of cancer incidence or mortality rates over long times, it is very important to use some standardized rates adjusted for the confounding of age or other factors. The age-standardized rate using the world standard population is widespread used. It is well known that there are two methods for estimating the age-standardized rates, e.g. directly and indirectly methods. The detailed descriptions on the methods for statistical analysis of cancer registration data can be seen in several reports[1,2,7] of IARC Scientific Publications. Here we will give a short review for the statistical methods related to calculation and comparison of adjusted rates in cancer registration in this section.

Calculation of Adjusted Rates and its Confidence Intervals

Let n_i be the number of cases in the *i*th age groups which diagnosed in a specified area, p_i be the number of population (frequently by sex), and a_i denote the age-specific rates for the *i*th age group. The age-standardized rates (ASR) for 18 age groups using the directly method can be written as

$$ASR = \frac{\sum_{i=1}^{18} a_i \bullet w_i}{\sum_{i=1}^{18} w_i},$$

where the w_i is the number of standard population (World) in the *i*th age group. Its variance under the Binomial approximation[1] can be estimated from

$$Var(ASR) = \frac{\sum_{i=1}^{18} [\, a_i \bullet w_i^2 \bullet (100000 - a_i) / p_i \,]}{(\sum_{i=1}^{18} w_i)^2}$$

If under the Poisson approximation[1] provided that the age-specific rates a_i are small, an alternative variance formula has the form of

$$Var(ASR) = \frac{\sum_{i=1}^{18} (a_i \bullet w_i^2 \bullet 100000 / p_i)}{(\sum_{i=1}^{18} w_i)^2}.$$

Thus the approximation (100-α)% confidence intervals for age-standardized rates can be expressed as

$$ASR + \_ \, Z_{\alpha/2} \times \sqrt{\hat{Var}(ASR)},$$

in which, $Z_{\alpha/2}$ is a standard normal quartile.

When the number of cases for some age groups are very small, or the age of cases is unknown the indirectly method can be used to calculating the age-standardized rates for cancer registration. First to calculate the expected number of cancer cases, then to calculate the standardized ratio. For cancer incidence analysis it is called the standardized incidence ratio (SIR), and for mortality the standardized mortality ratio (SMR). Suppose that a_{is} is the age-specific rates of the standardized population, let c_i be the expected number of cases in the corresponding *i*th age groups. The expected number of cases e_i can be estimated from

$$e_i = a_{is} \bullet p_i / 100000 \times a_{is} \times p_i .$$

Thus the standardized ratio can be calculated by dividing the observed number of cases by the expected number of cases (for example SIR)

$$SIR = \frac{\sum_{i=1}^{18} n_i}{\sum_{i=1}^{18} e_i} \times 100 .$$

The variance of standardized ratio can be given by

$$\hat{\text{Var}}(\text{SIR}) = \frac{\sum_{i=1}^{18} n_i}{(\sum_{i=1}^{18} e_i)^2} = \frac{\sum_{i=1}^{18} n_i}{(\sum_{i=1}^{18} a_{is} p_i / 100000)^2} ,$$

and the approximate (100-α)% confidence intervals for the standardized incidence ratio can be estimated as the same way of directly adjusted rates, e.g.

$$\text{SIR} +\_ [Z_{\alpha/2} \times \sqrt{\text{Var}(\text{SIR})}] ,$$

or the lower and upper limits for confidence[2] can be constructed as follow

$$\frac{\left[\sqrt{\sum_{i=1}^{18} n_i - Z_{\alpha/2} \times 0.5}\right]^2}{\sum_{i=1}^{18} e_i} - \frac{\left[\sqrt{\sum_{i=1}^{18} ni + Z_{\alpha/2} \times 0.5}\right]^2}{\sum_{i=1}^{18} e_i} .$$

Cumulative Rates and Risk

The cumulative rates[7] is an another age-standardized rate frequently used in cancer registration that is the sum over each year of age of the age-specific incidence rates taken from birth to age 64 or 74 (the 0–64 or 0–74 rate). The formulas are

$$\text{C-Rate}(0\text{–}64) = \sum_{i=1}^{13} 5 \times a_i \quad \text{and}$$

$$\text{C-Rate}(0\text{–}74) = \sum_{i=1}^{15} 5 \times a_i ,$$

for the 0–64 and 0–74 rates respectively. If the age classes used are 0-, 1-, 5-, 10-, ... then the cumulative rates are of the forms

$$\text{C-Rate}(0\text{–}64) = 1 \times a_1 + 4 \times a_2 + \sum_{i=3}^{13} 5 \times a_i ,$$

$$\text{C-Rate}(0\text{–}74) = 1 \times a_1 + 4 \times a_2 + \sum_{i=3}^{15} 5 \times a_i .$$

Under the Poisson approximation, the variance of cumulative rate can be expressed as

$$\text{Var}(\text{C-Rate}) = \sum_{i=1}^{18} \frac{t_i^2 \times a_i}{p_i} ,$$

where the t_i is the length of the age-intervals. For there are 18 age groups the all of the ti are equal to 5, and the 19 age groups the t_1=1, t_2=4 and others are equal to 5. The confidence intervals of cumulative rates can be constructed by using the same way as that of age-standardized rates described before. Using the cumulative rate, one can calculate the cumulative risk

$$\text{C-Rate} = 100 \times [1 - \exp(\text{C-Rate}/100)] .$$

Testing the Hypothesis of Adjusted Rates

Comparison of Age-standardized Rates (Directly Method)

For comparison of age-standardized rates, it is usual to calculate the standardized rate ratio which is the ratio between two directly age-standardized rates and test it whether the observed ratio is significantly different from unity. First to calculate the standardized rate ratio (ASR_1/ASR_2), then to calculate the confidence intervals of the standardized rate ratio, the third step is to test the hypothesis. If the confidence intervals (e.g. 95%) do not include 1.0, one can say at the 5% level the two standardized rates are significantly different. An approximation confidence intervals[8] of the ratio is

$$(ASR_1/ASR_2)^{1 \pm (Z_{\alpha/2}/X)} ,$$

in which

$$X = \frac{(ASR_1/ASR_2)}{\sqrt{\text{Var}(ASR_1) + \text{Var}(ASR_2)}} ,$$

and $Z\alpha/2$ is the standard normal quartiles, for example one can select $Z\alpha/2$=1.96 at the 95% significant level or $Z\alpha/2$= 2.58 at the 99% level.

Comparison of Age-standardized Rates (Indirectly Method)

As the same way of comparison of the directly adjusted rates, to calculate the appropriate confidence intervals and to see whether the intervals include or exclude the value of 100 for testing the significance of indirectly adjusted ratios. When the 95% (or 99%) confidence interval excludes the value of 100, then it can be concluded that the incidence rate of interest is significantly higher than that of rate of standard population selected by the investigator (at the 5% level of significance).

Testing the Linear Trend of Adjusted Rates

The relationship between the years (χ) that cases diagnosed and the age-standardized rates (y) during each year can be expressed as a standard linear regression model[2]

$$y = \alpha + \beta x$$

where α is the intercept, β is the coefficient of regression line. In practice, the years of cases diagnosis is often denoted by the numbers of some natural order (e.g. 1, 2, ...), then to estimate the regression coefficient of the linear model. The coefficient can be estimated from

$$\beta = \frac{\Sigma x_i\ y_i - \dfrac{\Sigma x_i\ \Sigma y_i}{n}}{\Sigma x_i^2 - \dfrac{(\Sigma x_i\ y_i)^2}{n}},$$

where n is the number of the observation years. The variance of β is given by

$$\mathrm{Var}(\beta) = \frac{\dfrac{1}{n-2}\left[\Sigma (y_i-\bar{y})^2 - \beta^2 \Sigma (x_i-\bar{x})^2\right]}{\Sigma (x_i-\bar{x})^2},$$

where

$$\bar{y} = \frac{\Sigma y}{n}, \qquad \bar{x} = \frac{\Sigma x}{n}.$$

Thus the statistic t is easily calculated from

$$t = \frac{\beta}{\sqrt{\mathrm{Var}(\beta)}},$$

and it follow a t-distribution with n-2 degrees of freedom. This statistic can be used to testing the significance of regression coefficients.

PROGRAM DESCRIPTION

The CANTEST is a menu-driven program which was written is BASIC language. It was designed to check automatically any input errors from keyboard by user, and will give some indications on the screen. The program can run on any IBM microcomputer or its compatible computer with the standard BASIC language (Version 3.20 or later). The executable file can be generated using the Microsoft BASIC compiler program and then run directly under Disk Operation System (DOS) . Table 1 is the input screen or menu of program CANTEST.

Table 1. Input menu of program CANTEST

C:\>CANTEST [Enter]
1. Testing for directly adjusted rates
2. Testing for indirectly adjusted rates
3. Testing for linear trend of adjusted rates
4. Confidence intervals for directly adjusted rates and cumulative rates (0–64 and 0–74)
9. Quit (to DOS)
Please selection (1–4, 9):1

The source program code have about 1000 lines including a main program and six subroutines which are for the data input or read, testing for directly adjusted rates, testing for indirectly adjusted rates, testing for linear trend of adjusted rates, calculating the confidence intervals for cumulative rates and risk, error checking messages. The program occupy about 40k hard disk space. No special requirement was needed for computer hardware.

There two styles of data input. The first is the data is read from keyboard step by step and the second

is from a disk file (ASCII file). The program can handle the data sets there are 18 or 19 age groups. The first line of data file is the number of age groups (18 or 19), years or periods of analysis, age-standardized rates (ASR). The other 18 or 19 lines of disk file are the age-specific rates, the corresponding number of population during the same year or period, number of World standard population. The columns of disk file are different according to the items of analysis. The age-specific and standardized rates can be obtained from our another program CANSTAT[9] or other computer package.

Table 2. Comparison of age-standardized rates for oesophageal cancer for males during 1974–1978 and 1984–1988 in Urban Shanghai

Age groups	1974–1984		1984–1988		World population
i-	a_i	p_i	a_i	p_i	w_i
0-	0.0	533396	0.0	1201788	12000
5-	0.0	652926	0.0	684380	10000
10-	0.0	1295876	0.0	638338	9000
15-	0.0	1935496	0.0	946342	9000
20-	0.1	1422882	0.0	1772897	8000
25-	0.2	1338166	0.2	24772897	8000
30-	0.3	891558	0.4	2048129	6000
35-	3.3	817243	1.2	1346615	6000
40-	6.6	984647	2.2	810005	6000
45-	17.1	1090815	4.1	868136	6000
50-	38.7	945887	12.6	1212914	5000
55-	65.3	753809	30.0	1049711	4000
60-	109.5	556960	64.0	835546	4000
65-	167.1	407561	86.6	612918	3000
70-	226.2	251958	127.3	408277	2000
75-	280.6	121176	154.4	226716	1000
80-	290.7	43005	190.5	95194	500
80+	261.4	11476	212.0	28958	500

Adjusted rates: $ASR_1 = 25.7/100000$ $ASR_2 = 13.6/100000$ Standardized rate ratio: $ASR_1/ASR_2 = 1.89$

95% confidence intervals: Binomial interval 1.78–2.00, Poisson interval 1.78–2.00

99% confidence intervals: Binomial interval 1.75–2.04, Poisson interval 1.75–2.04

Significance of difference: $P<0.01$ (at 1% level)

For each item of the program menu, the program will produce the corresponding confidence intervals or related statistics and the P value of significance. There are two results in output of our program, one is for Chinese and other is for English.

APPLICATIONS

Comparison of Directly Adjusted Rates

Table 2 is the output of our program for testing the directly adjusted rates of two periods. The example data sets are the oesophageal cancer incidence rates for male during 1974–1978 and 1984–1988 in Urban Shanghai. The standardized rate ratio is 1.89. Its 95% and 99% confidence intervals under Poisson and Binomial approximations was also listed in Table 2. The results indicated the difference between two directly adjusted rates of periods 1974–78 and 1984–88 for male oesophageal cancer had

highest significance (at 1% level).

Comparison of Indirectly Adjusted Rates

For calculating the age-standardized incidence ratio (SIR) of oesophageal cancer for male in Urban Shanghai in 1985, we select the age-specific incidence rates of 1972 as a standard rates. The age-specific rates of 1972, number of death and population of 1985 are listed in Table 3. The observed number of death is 504, and expected number of death is 1311.617. The standardized incidence ratio (SIR) is equal to 38.43. Its 95% and 99% confidence intervals are listed in Table which do not include the value of 100, so it can be concluded that the observed rate of oesophageal cancer for males in 1985 was significantly lowest than that of 1972 (at the 1% level of significance).

Table 3. Calculation and testing of indirectly adjusted rate of oesophageal cancer for males in Urban Shanghai in 1985 (using the rates of 1972 as standard)

Age groups	1972	1975	
i-	a_i	n_i	p_i
0-	0.0	0	253321
5-	0.0	0	135354
10-	0.0	0	119356
15-	0.0	0	177667
20-	1.1	0	333062
25-	0.6	1	494749
30-	2.2	1	421254
35-	5.6	4	284591
40-	13.4	4	163565
45-	27.5	6	159177
50-	49.6	28	240691
55-	91.9	52	211643
60-	130.7	116	168407
65-	229.6	77	123180
70-	255.5	81	82972
75-	280.6	75	46070
80-	222.7	47	19608
80+	203.5	12	6066

Observed death number: 504 Confidence intervals: 95% (460.96, 548.96); 99% (447.74, 563.58)
Expected death number: 1311.617
Standardized incidence ratio (SIR): 38.43 Confidence intervals*: 95% (35.14, 41.85); 99% (34.14, 42.97)
Significance of difference: P <0.01 (at 1% level) * The methods for calculating intervals see section 2

Testing the Linear Trend of Adjusted Rates

Table 4 is the age-standardized rates of oesophageal and gastric cancers diagnosed in Urban Shanghai during several time periods (from 1972–74 to 1987–89).[10] We used these rates to perform the statistical tests for linear trend of these rates. The results indicate there is evidence of a significant decrease in the incidence of oesophageal and stomach cancers in Urban Shanghai between 1972–74 and 1987–89 except for female stomach cancer. For male oesophageal cancer the age-standardized rate decreased by an average of approximately 3.25 cases per 100000 per period and 2.45 cases for male stomach cancer and 1.23 cases for female oesophageal cancer.

Table 4. Testing for linear trend of adjusted rates (1/100000) of oesophageal and gastric cancers in Urban Shanghai (from 1972–74 to 1987–89)

Time period	Order	Males		Females	
		Oesophageal	Stomach	Oesophageal	Stomach
1972–74	1	28.8	62.0	11.3	23.9
1975–77	2	24.7	59.2	10.4	24.8
1978–80	3	22.5	56.9	9.2	24.3
1981–83	4	17.1	54.5	7.7	22.4
1984–86	5	14.4	51.3	6.4	21.7
1987–89	6	13.3	50.1	5.4	23.2
Statistics:					
Intercept:		31.25	64.23	12.70	24.85
Slope of regression line:		-3.25	-2.45	-1.23	-0.42
Confidence intervals: 95%		(-3.79, -2.71)	(-2.66, -2.23)	(-1.31, -1.15)	(-0.88, 0.04)
99%		(-3.96, -2.54)	(2.73, -2.16)	(-1.34, -1.12)	(-1.03, 0.19)
T-value:		-11.7863	-22.1095	-289998	-1.7883
Degrees of freedom:		4	4	4	4
P-value:		0.000297	0.000025	0.000008	0.148251

Table 5. Calculation of confidence intervals of directly adjusted rates and cumulative rates (using the data 1974–78 in Table 2)

Confidence intervals for directly adjusted rates: ASR=25.7

Binomial approximation: 95% 24.97–26.43
99% 24.74–26.66
Poisson approximation: 95% 24.97–26.43
99% 24.74–26.66

Cumulative rates: 0–64 rate (%) 1.2055
Confidence intervals: 95% 1.1559–1.2551
99% 1.1403–1.2707
0–74 rate (%) 3.1720
Confidence intervals: 95% 3.0700–3.2740
99% 3.0378–3.3062

Cumulative risk: 0–64 (%) 1.1983
0–74 (%) 3.1222

Calculation the Confidence Intervals of Directly Adjusted Rates and Cumulative Rates

In section 4.2 we give an example for estimating the confidence intervals for indirectly adjusted rates (standardized incidence ratio SIR on standardized mortality ratio SMR) and testing the hypothesis or SIR or SMR. We do not provide the calculation of confidence intervals for directly adjusted rates in section 4.1 and give the example for comparison of two directly adjusted rates. Using the rates and numbers of oesophageal cancer for males during 1974–78 in Urban Shanghai in Table 2, we calculate the confidence intervals of age-standardized rate (25.7), cumulative rates (0–64 and 0–74) and its confidence intervals, and the corresponding cumulative risk. The results were listed in Table 5. The confidence intervals of directly adjusted rates of oesophageal cancer for males under Binomial approximation are very closer to the that of Poisson approximation.

PROGRAM AVAILABILITY

The copy of program CANTEST can be obtained by sent a 5-inch or 3.5-inch blank formatted diskette to the first author on request with no charge.

Acknowledgments

The development of program CANTEST was supported by Shanghai Cancer Registry. The authors thank all staffs of Shanghai Cancer Registry for the data sets

preparation.

REFERENCES

1. Maclennan R, Muir C, Steinitz R, et al. Cancer registration and its techniques (IARC Scientific Publications No.21). Lyon: International Agency for Research on Cancer. 1978.
2. Jensen OM, Parkin DM, Maclennan R, et al. Cancer registration: principles and methods (IARC Scientific Publications No. 95). Lyon: International Agency for Research on Cancer. 1991.
3. Coleman MP, Esteve J, Damiecki P, Arslan A, et al. Trends in cancer incidence and mortality (IARC Scientific Publications No. 121). Lyon: International Agency for Research on Cancer. 1993
4. Kupper LL, Janis JM, Karmous A, et al. Statistical age-period-cohort analysis: a review and critique. J Chron Dis 1985; 38:811.
5. Robertson C, Boyle P. Age, period and cohort models: the use of individual records. Stat Med 1986; 5:527.
6. Immonen-Raiha P, Hatonen S, Torppa J, et al. A statistical analysis system macro for age-standardized incidence rates. Computer Methods and Programs in Biomedicine 1994; 44:79.
7. Day NE. Cumulative rates and cumulative risk. In: Muir C, Waterhouse J, Mack T, et al. eds. Cancer Incidence in Five Continents, Vol . V (IARC Scientific Publication No. 88). Lyon: International Agency for Research on Cancer. 1987; p787–789.
8. Smith P. Comparison between registries: age-standardized rates. In: Muir C, Waterhouse J, Mack T, et al. eds. Cancer Incidence in Five Continents, Vol. V (IARC Scientific Publications No. 88). Lyon: International Agency for Research on Cancer. 1987; p790–795.
9. Xiang YB, Yuan JM, Gao YT, et al. A microcomputer program for statistical analysis of data from cancer registration. Tumor (Shanghai) 1993; 6(12):1.
10. Zheng W, Jin F, Devesa SS, et al. Declining incidence is greater for oesophageal than gastric cancer in Shanghai, People's Republic of China. Br J Cancer 1993; 68:978.