ACTA ONCOLOGICA

Vol. 29
Fasc. 8

1990

EPIDEMIOLOGICAL STUDIES OF GEOGRAPHIC VARIATIONS OF CANCER INCIDENCE IN SWEDEN

Choice of variables and statistical units

L. EHRENBERG, G. EKMAN and Å. SVENSSON

Abstract

The Swedish Cancer-Environment Registry has been used for studies of the parts of cancer incidences and of their variations between subpopulations, which, in a statistical sense, can be 'explained' by environmental variables. In previous studies the dependence of age-standardized incidences in municipalities (279 in Sweden) on population density, socio-economic variables, smoking habits and a variable for 'diagnostic intensity', assumed to allow for variations in under-diagnosing and under-reporting, was studied in, i.a., multiple-regression analyses. Due to intrinsic variation in potentially etiologic factors and in size, municipalities are less suitable as geographic units. Therefore a system was developed to build up, from parishes (about 2 600), pseudo-municipalities standardized with respect to size and some environmental variable(s), in the present study population density. This paper shows, for a number of cancer diseases and for total cancer, the superiority of pseudo-municipalities to municipalities with respect to power of explaining variation and, preliminarily, also incidence.

Key words: Cancer, incidence, Sweden, register study, geographic variations, environment, variables, statistical units.

In Sweden and the other Nordic countries the compulsory reporting of, in principle, all cases of cancer to central registries renders epidemiologic 'register studies' a promising source of information on the structure and variations of incidences of tumour diseases. For research purposes the Swedish 'Cancer-Environment Registry' (CER) (1) was established in 1978 by linking the Cancer Registry (for the years 1961–1973) with certain data from the 1960 population census, such as data on residence and employment. (The registry has later been updated, i.a. by introducing data from the 1970 population census.)

The mentioned linked registry has been used in a number of epidemiological studies which, besides development and testing of statistical models, aimed at estimates of the fractions of the incidences of tumour diseases, which, in a statistical sense, are 'explainable' in terms of environmental variables such as population density, socio-economic variables and smoking habits. ('Explainable', 'explained' etc. is given with quotation marks in order to indicate association in a statistical sense with independent variables that are mostly unable to identify causes of disease.) Certain results of these studies were utilized in the Report of the Swedish Committee on Cancer Prevention (2), with details published (in Swedish) in Appendices 1–4 of that Report (3–6). A preliminary study of the magnitude of the 'urban factor' has been carried out (7). In the mentioned studies, multiple-regression and Poisson models were applied to incidence data for the approximately 280 Swedish municipalities ('townships'), characterized by environmental variables selected by means of factor analysis from a large number of variables.

The idea of using registry data for studies of associations between cancer incidence and regional socio-economic data etc. is not new. A comprehensive investigation with such aims of variations of cancer incidence in Finland, with methods different from the present ones, has been carried out by Teppo et al. (8).

Sweden is geographically and administratively subdivided into 24 counties, consisting of municipalities (in all

From the Department of Radiobiology, and the Division of Mathematical Statistics, Stockholm University, Stockholm, Sweden.

Submitted 15 December 1989.
Accepted for publication 24 April 1990.

284 at present), consisting in turn of parishes (approximately 2 600, a time-honored concept; each parish centers about a (state) church). At present, due to the extensive computerization of all statistical data in Sweden, practically all kinds of statistics may be obtained on a parish level if requested. However, statistical publications usually report on a county or municipal level.

Epidemiological analyses of municipality data are complicated by the variation in population size of these administrative units (range in Sweden 3 000–600 000), which i.a. leads to a variable precision of certain descriptive variables, e.g. for smoking habits which is based on interviews of a random sample of 1% of the population of ages 18–69 (9, 10). Municipalities further often exhibit a great intrinsic heterogeneity with respect to factors such as population density, with a strong correlation with cancer incidence. In this latter respect parishes (average size approximately 150 km^2, 3 100 persons) are more homogeneous. Official statistical data with potential explanatory power are available on the parish level, but since the parish is too small a unit, a method was developed to build up, from parishes, a geographic unit, here called 'pseudo-municipality', which is standardized with respect to size and some important descriptive variable(s).

The present paper has the main purpose of comparing pseudo-municipalities with administrative muncipalities with respect to the explanatory power of mainly socio-economic variables in multiple-regression analyses of incidences of cancer diseases.

Material and Methods

Registries and database

The record unit of the CER (1) is a case of reported cancer. The unit contains certain information concerning the tumour (localization, histological type, etc). The unit also contains information concerning the person: date of birth, sex, domicile (parish), occupational status (standard classification codes for branch and position), education (rudimentary classification), place of work (parish designation).

Adjoined to the CER is a background registry giving the population (in 1960) for every intersection of age group, domicile parish, sex, education, occupational code, etc. Thus aggregates of cases and incidence computation for any given aggregate may be obtained.

The parish identification permits the introduction of a whole series of socio-economical and other variables tied to the parish. By projection these variables may be affixed to individual cases, and by summation or weighted mean values to aggregates of cases. Examples of such variables are: population age profiles, age–sex distribution, measures of population density, ('degree of urbanization'), and automobile density (cars per inhabitant). A number of 'socio-economic' variables based on occupational characteristics had been assembled by K. Berglund at The National Institute of Environmental Medicine (6).

Other variables have been adjoined, using more circuitous methods:

- gamma-radiation measures (data available for 55% of the parishes);
- smoking data (ratio of smokers to adult population, obtained from a large-scale sample survey from the 1960s (9, 10);
- measure of diagnostic sophistication tentatively constructed to allow for temporal and geographic variations in diagnostic and associated reporting procedures (cf. refs. 2, 3; this variable is described in the following section).

'*Diagnostic intensity*'. Each CER unit contains information about the diagnosis of the case with respect to diagnosing hospital, and often also clinic/physician, and a codification of the diagnostic method. Empirical evidence, epidemiological studies, our own preliminary and later on more comprehensive statistical analyses, indicated a strong, and most probably causal, correlation between diagnostic practice and reported incidence rates of various cancers (3, 5). The importance of this problem has been recognized repeatedly; cf. references given by Gori & Lynch (11) as well as the efforts of these authors to measure underreporting of cancer mortality. In the mentioned Finnish study, Teppo et al. (8) discussed the possible association of variations of the incidence of 'cancer at unspecified primary sites' (ICD 199) with diagnostic effectiveness.

The diagnostic-intensity variable is defined as the fraction of all cancer cases reported that were diagnosed by histological examination of surgical or biopsy materials (diagnosis code 3) or by autopsy with histological examination (code 4). For each geographic area considered the diagnostic intensity is then calculated as the weighted mean of the values for the hospitals serving the area.

Pseudo-municipality, a statistical unit for studies of incidence variations. In order to overcome the problems mentioned introductorily of referring incidence to administrative municipalities, a method was developed to construct a unit, 'pseudo-municipality', which is homogenized with respect to size and potential explanatory variables. Since demographic and socio-economic data can be linked to parishes, the only geographic unit in Sweden which combines continuity with reasonable size, aggregation has to proceed from parishes. The requirement for homogenization entails some kind of nearness, if not contiguity, of parishes to be included in a pseudo-municipality.

As yet it has not been possible to describe each parish in terms of a central coordinate and an area. There does not seem to exist any practical algorithm for determining an

optimal grouping of parishes into pseudo-municipalities, given the number of municipalities and/or a target size with permissible deviations as input. To solve this problem we have employed a short-cut method, using the official numbering of the parishes which contains certain contiguity information. In principle the procedure is as follows: The units were intrinsically homogenized with respect to population density, a description exhibiting strong variations between parishes within administrative municipalities (7). This description is purposely brief; a detailed description is of greater interest to the computer technician, and could not be included in the limited scope of this presentation. The computer programs used for this construction may be made available to interested parties.

(a) all parishes exceeding the upper limit of the target size are classified as pseudo-municipalities and removed from further consideration;
(b) starting at one end of the country (south or north) the largest parish of a geographically extreme municipality is chosen as a node;
(c) using the contiguity information of the numbering, and taking into consideration supplementary homogeneity conditions (in this study: population density (12)), parishes are added to the node until the target size is exceeded;
(d) when a municipality is exhausted, i.e. all parishes fulfilling the conditions preset are used, the procedure is continued into neighbouring municipalities;
(e) a first 'pass' of the above kind is followed by a second adjustment phase, in which certain (usually small) parishes are 'exchanged' when obvious contiguity properties can be improved.

The results described in the following are based on a target size = the average municipality size in Sweden (the number of statistical units are thus identical in comparisons between the two approaches). In exploratory studies of the influence of unit size which varied from 15 000 and upwards, the highest values of multiple correlation coefficients were found at approximately 25 000 inhabitants. In the present study the average size is about 32 000, i.e. not far from the optimum.

Analyses

The general analytical approach was as follows:

– Age-specific incidences of some 20, partly broad, groups of cancer diseases were calculated for each municipality or pseudo-municipality. Enclosed in the study were persons of age of 20 years or more in 1960. Standardization of the incidences in 1961–1967 to the age distribution of the 1960 Population Census was done for age cohorts born in the intervals 1931–1940, 1921–1930 . . .1881–90, ⩽1880.

– From the pool of candidate variables a set of descriptor variables was selected by means of factor analysis, with the superimposed conditions that variables (such as smoking data), which may in some way be interpreted in terms of causative factors, should be included and further that population density (as an expression of degree of urbanization) and car density should be included. For the purpose of comparing municipalities and pseudo-municipalities the present paper describes analyses with the same set of variables as was used in earlier studies of municipality incidences although this set is somewhat suboptimal in the pseudo-municipality approach. In Table 1 the studied descriptive variables (1–8) are listed, together with data for their variation, their interdependencies, and notes on what they are thought to express. It should be noted (cf. interdependencies in the last column of Table 1) that the variables mainly express various traits of urbanization. Due to strong negative correlation variables 2 (population density) and 4 (for rural character) are nearly tautologic, and variable 4 was therefore excluded in the main analyses.

– Multiple regression analyses were carried out with the age-standardized incidence rate for each tumour site or site group as the dependent variable.

– An examination of significances of the (partial) regression coefficients indicates which descriptors that have explanatory power.

Besides being limited to the sites or groups of sites which were included in the earlier study of municipality data (Table 2) this presentation is restricted to incidences in the years 1961–1967, because of sufficient nearness in time to the 1960 census data. In the results here reported, 5 pseudo-municipalities formed from the parishes of the city of Malmö, served by the General Hospital of Malmö, and the pseudo-municipalities served by the Karolinska Hospital in the Stockholm area were excluded. The reason for this is the extreme values of the diagnostic intensity in these areas, particularly in Malmö (13, 14), with the possibility that these few values could prevent the detection of general relationships prevailing in the country. Parallel analyses with these areas included showed practically the same results, although with a strengthened influence of the diagnosis variable.

Results

For sites studied with both geographic units the explainable parts of the total variations (variances), estimated by the (squared) multiple regression coefficients, are given in Table 2. Significances of regression coefficients are given in Tables 3a and b for males and females respectively. These analyses were carried out with the 7 variables Nos. 1–3 and 5–8 in Table 1. The influence of introducing variable 4 on regressions for colorectal cancer and breast cancer are presented in Table 4.

Table 1

Variables used in the study of cancer incidences of Swedish municipalities

Variable	Coefficient of variation	The variable expresses	Correlations (positive and negative) with other variables[a]
1 Diagnostic intensity	0.064	First, influence of diagnostic routines on reported incidence; maybe also large city	Var. 6 (pos.) Var. 3 (neg.)
2 Degree of urbanization (population density)	0.376	Air pollutants? life-style factors including smoking and alcohol habits	Var. 4 (strongly neg.) Var. 5, 6, 7, 8, (pos.)
3 Cars per inhabitant	0.237	Certainly not motor exhausts (sic!). Socio-economic factors, expecially manufacturing industry (see last column)	Var. 5 (strongly pos.) Var. 4, 7 (pos.) Var. 1 (neg.) Not var. 2
Fraction of population employed in			
4 agriculture and forestry (men)	1.067	Rural, or sparsely developed areas; also absence of urban character	Var. 2 (strongly neg.) Var. 5, 6, 7, 8 (neg.)
5 manufacturing industry and mining (men)	0.394	Factory town/villages	Var. 2, 3 (pos.) Var. 4, 6, 7 (neg.-strongly neg.)
6 bank and insurance (men)	0.899	Metropolitan and large towns	Var. 1, 2 (pos.) Var. 4, 5 (neg.)
7 public work (women)	0.366	County centers; hospital towns	Var. 2, 3 (pos.) Var. 5 (strongly neg.) Var. 4 (neg.)
8 Fraction habitual smokers (men and women)	0.277	Smoking habits and correlated life-style traits (highest in large towns)	Var. 2 (pos.)

[a] Partial correlations, i.e. correlations that remain after elimination of the influences of the other variables, in the analysis of municipalities (3).

Estimates of the parts of the incidences which seem to be 'explained' by the multiple-regression model and the variables applied, are presented in Table 5. These estimates are preliminary.

The comparison has been made in the framework of a linear regression model for the age-standardized incidences. There are of course other possible models, such as Poisson regression. Preliminary results obtained by applying a Poisson model lend support to the present results; they will be discussed in a forthcoming paper.

Discussion

'Explained' part of total variation

The simplest way of comparing the multiple regression analyses based on pseudo-municipalities and on administrative municipalities is to consider the 'explained' parts of variation, which is conveniently expressed by the squares, r^2, of the multiple correlation coefficients, r . r^2 is equal to the 'explained' part of the total variance.

As shown in Table 2 a shift from municipalities to pseudo-municipalities leads in most cases to higher values of r^2, i.e. increased explainable parts of the variances. This is the expected consequence of shifting part of the variation from within units to between units. A contribution to the increase of r^2 is further obtained through the size-homogenization of the units: due to the great variation in size of the municipalities a large number of small units cause a greater random—and therefore unexplainable—variability than in the analyses based on pseudo-municipalities. (It should be remarked that in the previous analyses of municipalities, given in Table 2, incidences were not weighted with respect to size of the units.)

Table 2

'Explained' parts (r^2) of total variances of incidence of (groups of) tumour diseases in municipalities and pseudo-municipalities (for definition see page 962)

Site	ICD 7	Sex	Average number of cases per unit	Explained part of total variance r^2	
				Municipality	Pseudo-municipality
Oesophagus	150	M	4.1	0.18	0.32
		F	2.3	0.07	0.17
Stomach	151	M	39.8	0.07	0.10
		F	24.6	0.01[a]	0.09
Colon and rectum	153–154	M	40.3	0.14	0.32
		F	36.6	0.12	0.24
Pancreas	157	M	12.1	0.07	0.21
		F	9.8	0.04	0.12
Lung, etc.	162	M	29.1	0.62	0.76
		F	7.8	0.06	0.30
Breast	170	F	77.6	0.20	0.36
Cervix uteri	171	F	54.7	0.42	0.44
Corpus uteri	172–174	F	23.8	0.06	0.14
Prostate		M	57.4	0.30	0.29
Kidney		M	17.6	0.39	0.43
		F	9.3	0.30	0.29
Urinary bladder	181	M	15.9	0.28	0.44
		F	6.5	0.13	0.28
Leukemia	204–205	M	10.8	0.03	0.02
		F	8.3	0.03	0.02
All cancer	140–209	M	303.0	0.52	0.74
		F	355.0	0.42	0.60

[a] Not significant (>0.05).

Table 3a

*Significant partial regression coefficients for age-standardized incidences in municipalities/pseudo-municipalities (see page 962) in analysis with 7 descriptive variables (1–3, 5–8 in Table 1). Cancer incidences for men in the period 1961–1967. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. (*) p approaching the limit. Minus sign, partial regression negative*

Site	Variable No.						
	Diagnostic intensity	Population density	Car density	Percentage employed in			Regular smokers %
				Manufacturing, etc.	Bank and insurance	Public service	
	(1)	(2)	(3)	(5)	(6)	(7)	(8)
Oesophagus		/*	−**/−**		**/*	−*(*)/	*(*)/
Stomach	/*		−**/		/−**	/*	
Colon and rectum	*(*)/**		**/−**	*(*)/	**/**		
Pancreas, etc.			/−**		/**	/*	*/
Lung, etc.	**(*)/***	**(*)/(*)	−**/−***		***/***	*/	*(*)/***
Prostate	*/***	−*/−**	−**/−**	**/(*)	***/**(*)	**/	/**
Kidney	**/**		/−***	/**	**/**		
Urinary bladder	**(*)/***		/−**		**/**		
Leukemias	−(*)/						
All cancer	***/***		/−***	**/	***/***	*(*)/	**/**

Table 3b

*Significant partial regression coefficients for age-standardized incidences in municipalities/pseudo-municipalities (see page 962) in analysis with 7 descriptive variables (1–3, 5–8 in Table 1). Cancer incidences for women in the period 1961–1967. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. (*) p approaching the limit. Minus sign, partial regression negative*

Site	Variable No.						
	Diagnositc intensity	Population density	Car density	Percentage employed in			Regular smokers %
				Manufacturing, etc.	Bank and insurance	Public service	
	(1)	(2)	(3)	(5)	(6)	(7)	(8)
Oesophagus	/−*		/−***		/−*	−(*)/	/***
Stomach	/***				/−*		
Colon and rectum	*/**		(*)/	*/(*)	**/		−*/
Pancreas	−*/−(*)		/−***				/**
Lung, etc.	/*(*)		/−**		/**		/(*)
Breast	/**	−*/	/−**	***/**	***/***	(*)/	/−*
Cervix uteri[a]	**(*)/**	**/*	/−***			*/	**/
Corpus uteri		**/	/(*)				
Kidney	/*		/−**				(*)/
Urinary bladder	**(*)/*(*)				**/***	−*/	/−*
Leukemia		/−*		/*	/**		−*/
All cancer	***/***		*/−***	**/**	***/***	*/	

[a] Including carcinoma in situ.

Table 4

Influence of the introduction of the variable for rural character (No. 4) on pattern of significant partial regression coefficients for incidences in pseudo-municipalities of cancers of colon and rectum and of breast. For significances see Tables 3a and b

Site		Variable No.								r^2
		Diagnostic intensity	Population density	Car density	Percentage employed in			Regular smokers %	Employed in agriculture	
					Manufacturing, etc.	Bank and insurance	Public service			
		(1)	(2)	(3)	(5)	(6)	(7)	(8)	(4)	
Colon and rectum	M	**		−**		**			− − −	0.32
		**	**	−***		*			***	0.36
	F	**			(*)				− − −	0.24
		**	**	−**				−*	***	0.28
Breast	F	**		−**	**	***		−*	− − −	0.36
		**	**	−***	*(*)	***		−**	**(*)	0.40

Patterns of correlation

It appears from Tables 3a and b that by and large the observed pattern of variations in the pseudo-municipality approach is similar to the one obtained for incidences based on municipality (3, 6), a fact that strengthens the reliability of the analytical result. Certain differences are observed, however, and deserve being commented upon.

For most sites or site groups the diagnostic intensity (variable 1) exhibits positive correlations that are often highly significant and more so than in the previous analysis on municipality basis. This difference between the two approaches is partly caused by the subdivision of large, highly urbanized administrative municipalities, where the diagnostic intensity is highest, into several units and, conversely, a pooling of small rural units into larger ones. (The omission, for the purpose of avoiding an exaggerated influence of variable 1, of the largest and third-largest cities of Sweden, causes at the same time a certain underestimation of significances.) Although correlations of incidence with variable 1 certainly contain the bias, the estimation of which and correction for which was originally aimed at (3, 5), it is by all means obvious that it is also an expression of components of urbanization that are important to cancer risk. Variable 6, an expression of large-city character which is strongly correlated with variable 1, has evi-

Table 5

'Explainable' parts, $(\bar{I} - I'\ min)/\bar{I}$ of mean incidences ($\bar{I}$). I' min is the minimum incidence expected from the multiple regression model (cf. text 968).

Site		Incidence ($\bar{I}$) per 10^6	I'_{min}	'Explained' part of incidence: Present study (%)	'Explained' part of incidence: Previous estimate based on municipality[3] (%)
Oesophagus	M	56.8	16.5	71	50
	F	32.3	4.5	86	20
Stomach	M	554	370	33	22
	F	328	342	26	12
Colon and rectum	M	567	413(367[a])	35[a]	39
	F	493	323(273[a])	45[a]	24
Pancreas	M	170	112	34	35
	F	135	97	28	17
Lung, etc.	M	380	36	91	90
	F	90	46	48	48
Breast	F	1 044	803(770[a])	26[a]	23
Cervix uteri[b]	F	693[b]	333	52	50
Corpus uteri	F	331	241	27	20
Prostate	M	797	562	29	30
Kidney	M	180	102	43	39
	F	126	90	29	30
Urinary bladder	M	225	110	51	54
	F	85	24	72	41
Leukemias	M	135	116	14	20
	F	94	72	23	18
All cancer	M	4 266	3 042	29	30
	F	4 798	3 650	24	20

[a] Variable 4 included in analysis.
[b] Carcinoma in situ included.

dently been unable to counteract this shift to variable 1 of 'explaining' power related to high degree of urbanization.

Among the variables for urban character, No. 6, the proportion empolyed in bank and insurance, has the strongest 'explanatory' power, with positive correlations for most sites among men and for certain sites among women. A negative correlation with stomach cancer is indicated for both sexes. Variable 2, population density, has little to add, and the same may be said about variables 5 and 7, characterizing manufacturing industry center and district center respectively. With these variables several correlations obtained with municipality incidences disappear or are weakened after switching to pseudo-municipalities. On the other hand, the influence of the smoking variable (No. 8) is stronger when the incidences are expressed per pseudo-municipality. This is certainly to some extent a consequence of this explanatory variable becoming more reliable when it is based on a size-homogenized geographical unit.

Variable 3, the number of cars per inhabitant, would off-hand be understood to be a descriptor of urban conditions, related to the level of air pollutants. This is, however, not the case, this variable being positively correlated with rural conditions and particularly with employment in manufacturing industry (see Table 1). For this reason it is not astonishing that incidences are in most cases negatively correlated with this variable. These correlations are strengthened by changing to pseudo-municipalities. The car-density variable has evidently a large explanatory power and it is therefore important to clarify which socio-economic conditions it signifies. It is intuitively felt that this variable in some way is associated with a predominant influence of a 'healthy-worker effect'.

In the previous analysis of municipality data, colorectal and breast cancer exhibited a positive correlation with both the large-city variable (No. 6) and with the variable 4 for rural character, viz. proportion employed in agriculture and forestry. Since the change to pseudo-municipalities leads to a reduction of the number of rural units and an increase of the number of urban units contributing to pattern, it is of interest to see if the bi-centered pattern is retained. It is seen from Table 4 that this is the case. In a regression analysis without variable 4 the population density (variable 2) has no explanatory power since the residuals, after subtracting the influence from the other variables, tend to be high both for units with low and with

high population density. Using both the negatively correlated variables 2 and 4 in the analysis makes it possible to describe this variation. For both variables significant and positive regression coefficients are calculated. A similar pattern is found for cancer of oesophagus (men), lung (both sexes), and all sites (both sexes), and, with a tendency, for pancreas (men), and leukemia (women). No such effect was observed for the other sites studied.

'Explained' parts of incidences

A rough estimate of the 'explained' parts of the incidences may be obtained by a calculation of $(\bar{I} - I'_{min})/\bar{I}$, where $\bar{I}$ is the average incidence and I'_{min} the minimum incidence predicted by the model (3, 7). In previous analyses I'_{min} was estimated by the mean value of the two lowest predicted values and for the present comparison of the two geographical units 'explainable' parts of incidences were calculated in this way (Table 5).

As shown in Table 5, the 'explained' parts of the incidences are, in the present analysis, based on pseudo-municipalities equal to or larger than the values obtained in the previous study with municipality as geographic unit. The values are uncertain since the estimate presumes linear dependence on the variables applied. From the point of view that comparison is made with the minimum at an existing combination of variable values it is likely that a lower I'_{min} could be reached and that therefore the estimates are conservative. For instance, since smoking can be completely eliminated, at least in theory, the 'explainable' parts of incidences with a positive dependence on the smoking variable (Tables 3a, b) would be increased by additional 5–10% (in the case of cervix cancer more) if the value of this variable is set to 0% regular smokers. A certain underestimation, particularly of diseases associated with urbanization and smoking, is caused by the omission of certain highly urbanized areas. The values for all cancers are further underestimates because different tumours depend differently on the explanatory variables. This problem is being studied further and will be treated within the frame of a forthcoming comparative discussion of statistical models.

The low ability of the present approach to 'explain' incidences, and their variations, of stomach cancer and leukemias is noteworthy. Evidently, neither of these cancer types exhibit strong etiologic associations with the urban factor, smoking habits included (see also Tables 3a, b). In Sweden the incidence of stomach cancer is highest in the northernmost counties, and a stronger correlation might therefore have been obtained if a variable for latitude had been added. With respect to leukemias, completely different variables, including radiation and certain chemical exposures, would have had to be included to raise the explanatory power.

Above all, the fact that the descriptive variables are mean values and that therefore the true causative factors associated with the descriptors certainly exhibit a considerable variation within the units, the 'explainable' and in principle preventable parts of cancer incidences will always be underestimated in macroepidemiological studies of this kind.

Cause–effect relationships

Since the descriptive variables available for macroepidemiological studies of the kind reported here are only indirect expressions of causative factors, cause–effect relationships can only be discussed with utmost care. Care is also called for by the partly strong interdependencies of descriptors and by the fact that an intrinsic variation in the true causes remains. For these reasons the variable for diagnostic intensity and that for smoking habits can only partly be taken as direct measures of underdiagnosis/underreporting and influence of smoking, respectively, but should rather be seen as expressions among others of life-style habits.

Remembering these difficulties it is of interest to note the strong association of smoking habits with the incidence of lung cancer in men (Table 3a) obtained despite of the covariaton of smoking with urbanization variables, particularly population density (Table 1). Also incidences of all cancer and of prostate cancer in men and oesophagus and pancreas in women exhibit significant positive associations with the smoking variable.

The 'explainable' parts of the incidences studied seem to a large extent to reflect a positive association with various traits of the urban factor in a broad sense. This concerns particularly the variable '% employed in bank and insurance', which evidently expresses big-city character, but also the negative dependence on 'car density' (cars per inhabitant) should probably be interpreted in similar terms. It would be of interest to analyse further the nature of this effective descriptor.

In the quoted Finnish study (8), the incidence of lung cancer was relatively weakly correlated with socio-economic variables associated with urbanization, probably due to the advanced smoking habits also in rural areas of Finland. The strong influence of socio-economic variables for other cancers in the Finnish as well as in the present Swedish study supports the idea that the urban factor to a large extent has to be explained by living habits other than smoking, such as dietary habits. Urban air pollutants probably play a relatively small (although certainly not negligible) role, according to estimates based on exposure data (15, 16).

The bimodal association of the incidences of certain tumours with both rural and urban factors, observed when the variable '% employed in agriculture and forestry' is included, might, in the case of intestinal and breast cancer,

reflect influences of dietary habits such as high fat consumption which, as can be inferred from the official Family Expenditive Surveys (17), characterizes both big cities and certain agricultural areas. The observation of similar bimodal interdependences for cancer of the lung and a few other sites might indicate similar causal relationships and calls for further studies.

General conclusions and prognosis

The consistency, with previous studies and with expectation, of the results of the present study indicates that the introduction of the pseudo-municipality as the geographic unit for calculation of incidences in databased macro-epidemiologic investigations constitutes an improvement as compared with the corresponding use of administrative municipalities. A suitable starting point has thus been created for registry analyses, e.g. aiming at

- generation of hypotheses with respect to cause–effect relationships (2, 3);
- estimation of the portion of the national incidence that is 'explainable', in a statistical sense, by single or groups of environmental factors (2, 3, 7);
- identification of the components of and estimation of the size of the 'urban factor';
- calculation of expected numbers of cases (i.e. use as a relevant control) in studies of specific populations defined with respect to e.g. dietary habits or occupational exposures; especially this approach will facilitate relevant significance estimation in epidemiological surveillance systems;
- use as a basis for sampling.

For the interpretation of the observed patterns of partial correlations deeper studies, particularly of the nature of the great influence of variable 1, diagnostic intensity, are required. It is also expected that a subdivision of certain site groups may give a more clear-cut picture, despite the increased variances.

ACKNOWLEDGMENTS

This work was supported financially by the National Swedish Environment Protection Board, the Swedish Cancer Society and the Bank of Sweden Tercentenary Foundation. Previous studies of municipality data were sponsored by the Work Environment Fund.

Request for reprints: Dr L. Ehrenberg, Department of Radiobiology, Stockholm University, S-106 91 Stockholm, Sweden.

REFERENCES

1. National Board of Health and Welfare. The Swedish Cancer-Environment Registry 1961–1973. Stockholm, 1980.
2. Governmental Committee on Cancer Prevention, 'Cancer, Causes, Prevention, etc.', SOU 1984:67, Stockholm (in Swedish, English version to appear in 1991, Taylor and Francis, London). 1984.
3. Ehrenberg L, Ekman G. Cancer diseases in Sweden: Studies of variations in mortality and incidence (in Swedish). Ds S 1984:4, Appendix 1 to the Report of the Governmental Committee on Cancer Prevention, Stockholm, 1984.
4. von Bahr B, Bolander A-M, Ehrenberg L. Cancer mortality as a function of age, marital status and residence region, studied by means of a weighted multiplicative Poisson model (in Swedish). Ds S 1984:4, Appendix 2 to the Report of the Governmental Committee on Cancer Prevention, Stockholm, 1984.
5. Ekman G. Ehrenberg L, von Bahr B. Diagnostic intensity, a confounding factor in analyses of the Swedish Cancer-Environment Registry (in Swedish). Ds S 1984:4, Appendix 3 to the Report of the Governmental Committee on Cancer Prevention, Stockholm, 1984.
6. Berglund K, Ekman G, Dzieciaszek J. Influence of so-called socio-ecological variables on cancer incidence in the municipalities of Sweden (in Swedish). Ds S 1984:4, Appendix 4 to the Report of the Governmental Committee on Cancer Prevention, Stockholm, 1984.
7. Ehrenberg L, von Bahr B, Ekman G. Register analysis of measures of urbanization and cancer incidence in Sweden. Environ Int 1985; 11: 393–9.
8. Teppo L, Pukkala E, Hakama M, Hakulinen T, Herva A, Saxén E. Way of life and cancer incidence in Finland, A municipality-based ecological analysis. Scand J Soc Med Suppl. 1980; (Suppl 19): 84pp.
9. Statistics Sweden. Smoking habits in Sweden, a mail survey—Spring 1963. Stockholm: National Central Bureau of Statistics, 1965.
10. Cederlöf R, Friberg L, Hrubec Z, Lorich U. The relationship of smoking and some social covariables to mortality and cancer morbidity, Parts 1–2. Dept of Environmental Hygiene. Stockholm: Karolinska Institute, 1975.
11. Gori GB, Lynch CL. Decline of U.S. cancer mortality rates: Expert estimates of past underreporting. Regulatory toxicol Pharmacol 1986; 6: 261–73.
12. Statistics Sweden. Definition of 'degree of urbanization'. In: Population and Housing Census 1975, Part 3:1. Population in communes and parishes. Stockholm: Official Statistics of Sweden, 1977.
13. National Board of Health and Welfare. Cancer Incidence in Sweden 1959–1965. Stockholm, 1971.
14. Saxén EA. Trends: Facts or fallacy. In: Magnus K, ed. Trends in cancer incidence. Washington: Hemisphere Publ Corp, 1982.
15. Törnqvist M, Ehrenberg L. Approaches to risk assessment of automotive engine exhausts. IARC Sci Publ 1990; 104: 277–87.
16. Törnqvist M. Monitoring and cancer risk assessment of carcinogens, particularly alkenes in urban air. (Dissertation.) University of Stockholm, Stockholm, 1989.
17. Statistics Sweden. The Family Expenditure Survey 1969. Preliminary Results. Stockholm Stat Rep P 1971: 9.