A role for the ecological study in the developing world

F. SITAS, M. L. THOMPSON

Abstract **Retrospective case-control or prospective (follow-up) studies are important epidemiological tools and have provided useful information on exposure disease associations. Prospective studies would be the ideal option, but many countries (particularly in the developing world) do not have the necessary infrastructure to follow people up. Both retrospective and prospective studies are, however, sometimes conducted without due regard for their own limitations. These limitations are exacerbated when measures of exposure or disease are based on a single measurement and where the population under study is homogeneous with regard to exposure. The former is responsible for regression dilution bias and the latter for a lack of contrasts between exposure groups. Both factors would attenuate any relationship between exposure and disease. Ecological studies in epidemiology are weaker in design than case-control or prospective studies, but in some circumstances an ecological approach, which looks at the prevalence of an exposure or disorder in a number of areas of varying disease rates, may offer some advantages.**

S Afr Med J 1993; **83:** 753-756.

Epidemiological research in developing countries, and especially in South Africa, has to date largely been of a descriptive nature. If disease-exposure relationships are to be explored in a more analytical fashion, it is necessary to evaluate the extent to which different epidemiological study designs are feasible in the South African (and developing world) context.

Randomised controlled trials are the 'gold standard' in the establishment of disease-exposure associations. Many exposures of public health interest cannot, of course, be randomly allocated to subjects and must instead be investigated by means of observational studies. Prospective study designs are problematic in a developing world setting in that the necessary infrastructure for following up subjects over time is often absent. Retrospective case-control or cross-sectional methods are more practicable, and case-control studies are particularly effective when the disease under consideration is rare. Two potential limitations of observational studies are, however, often overlooked. These are measurement error (which leads to regression dilution bias) and lack of contrasts between cases and controls which results in poor heterogeneity. Both these factors will attenuate any association between disease and exposure.

A study design which may offer advantages in exploring disease-exposure relationships is the ecological study. If carefully executed and interpreted, ecological studies of varying disease and exposure rates across population subgroups may circumvent the problem of measurement error and lack of contrasts between exposure groups, and therefore generate useful hypotheses about disease-exposure associations. For this reason the ecological study may be useful to epidemiological researcher in South Africa and other developing countries.

Ecological studies

Ecological studies (geographical or aggregate studies) use a group of people as a unit of observation for disease or exposure. These units of measurement might be municipal districts, whole country populations, classrooms or factories, etc. Exposure is often measured by some summary index, for example mean number of cigarettes smoked, food consumption or gross domestic product.

An obvious limitation of ecological studies is the 'ecological fallacy', i.e. the possibility that aggregated variables might not reflect the status of individuals. Because of reliance on indirect measures of both explanatory and confounding variables, ecological studies may produce data of questionable validity from which inappropriate inferences could be drawn. This is especially so if confounders or effect modifiers are not taken into account. The literature on the ecological fallacy is extensive.[1-3] In certain circumstances, irrespective of statistical adjustment for potential confounders, associations that are quite the reverse of those expected can be obtained.[4] However, if carefully executed and interpreted, ecological studies of varying disease and exposure rates across population subgroups may generate useful hypotheses about disease-exposure associations. It seems reasonable to anticipate that the distribution of a disease should coincide with that of a putative causal agent.

Regression dilution

Any variable measured on a single occasion is prone to measurement error. This is because in a set of single measurements, persons with extremely high or low values are more likely to have been observed on a better or a worse than average day; this gives a distribution with large variance. The combination of further sources of error (machine, observer, laboratory) results in a distribution with still greater variance. For instance, in a study of 194 Boston women, Willett[5] found with regard to dietary interview data based on 1 day of intake, that those women in the top 90th percentile (for that day) consumed 3 times as much fat and 6,4 times as much vitamin A as those in the bottom 10th percentile. When the same women were observed over 4 weeks, the 10th to 90th percentile ratios (for the 4-week average) were reduced to 1,9 for fat and 2,5 for vitamin A.

Because of this variability in individual biological or exposure levels on the basis of single measurements on individuals, an association between exposure and disease will be attenuated, hence the term 'regression dilution'.[6,7] McMahon *et al.*,[8] for example, re-analysed data from 9 prospective studies on the relationship between blood pressure and coronary heart disease. After correction for regression dilution, the association was 60% greater than was originally found in the uncorrected analyses. Similar examples can be found in the single measurement of lung function,[7] cholesterol,[9] and to varying degrees, in most biological variables measured on a single occasion.

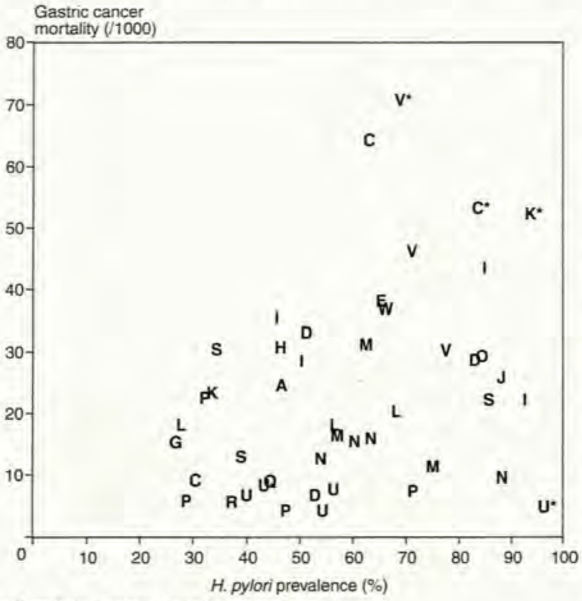
National Cancer Registry, Department of Tropical Diseases, South African Institute for Medical Research and University of the Witwatersrand, Johannesburg
F. SITAS, B.SC., M.SC. (MED.), M.SC. (EPIDEMIOL.), D. PHIL.
Department of Statistical Sciences, University of Cape Town
M. L. THOMPSON, B.SC. HONS, PH.D.

Accepted 5 Oct 1992.

In order to reduce regression dilution induced by measurement error, a measure of the 'within-person variability' is needed, often in the form of a validation sub-study.[5,8,10] It is therefore worth while when planning studies to ensure that a sub-sample or all study subjects are remeasured. However, to do this, follow-up of the subjects is required. This could be a difficult task, particularly in developing countries.

Ecological studies can to some extent circumvent regression dilution bias. If one considers a number of subgroups of a population (or a number of populations), each subgroup with its own mean level of exposure and its own disease incidence rate, then the average measure of exposure for a random sample from a particular subgroup will have the same mean value as the individual measures, but will have reduced variability. It can be shown (see Appendix) that, in the presence of measurement error, a regression of subgroup averages leads to a regression coefficient which is closer to the true coefficient (i.e. that without measurement error) than that based on individual measurements (which are subject to the bias induced by measurement error).

Asymptotically (i.e. for large samples within each subgroup) the regression coefficient based on subgroup averages approaches the true (without measurement error) coefficient. The same is not, however, true of correlation coefficients which can be artificially inflated (relative to the true correlation between measures on individuals) by averaging.



* Cumulative mortality up to 64 years of age.
Letters in scatterplot are the initials of each county.
R = 0,34; 2P = 0,2.

FIG. 1.
Correlation of gastric cancer mortality* and *Helicobacter pylori* antibody prevalence in 46 Chinese counties.

Heterogeneity

It is an almost knee-jerk epidemiological response to study a population where the problem is at its worst. However, by restriction of a study to one geographical area, there is often a good chance of discovering that the 'non-exposed' are actually 'exposed' but do not know about it. This poor heterogeneity of exposure between those 'exposed' and 'non-exposed' also results in an attenuation of the relative risk.[11] A greater sample size will also be required for the necessary power to detect differences in disease rates between those at only slightly different levels of exposure.[12] For example, most studies in industrialised countries have shown very weak (if any) associations between the risk of breast cancer and saturated fat intake. One of the reasons may be that fat consumption in industrialised countries is so ubiquitous[3] and dietary recall of fat intake is measured with such a high degree of error (cf. regression dilution bias) that the likelihood of finding a wide enough range of exposures and disease risk is rather low. Adjustment for regression dilution in the study by Willett *et al.*[5] did not make much difference to the relative risk between fat consumption and breast cancer.[10] It was only when international data from countries whose fat consumption varied were pooled that a potential association became more evident.[13] In fact, research on many chronic diseases within restricted study populations reaches a dead-end, and comparison of international data or data from populations in transition from one standard of living to another often provides clues for further research.[14,15] By incorporating the range of disease prevalences and exposure levels present in different areas, ecological studies may offer insights into potential disease-exposure relationships that would not be possible in studies in a single area with poor heterogeneity (i.e. with limited exposure range).

Choice of study design

If a common disease is the focus of investigation, then a number of regional substudies (looking at case-control or cross-sectional samples of individuals within each region) might suffice to reduce regression dilution and introduce heterogeneity of disease and exposure patterns, as was done, for example, in the American six-city air pollution study.[16]

A case-control study design is particularly suited to the study of rare diseases, If, however, a disease's incidence varies geographically it would also be amenable to an ecological study. Regional disease incidence figures (e.g. from official sources or sentinel surveillance centres) could be linked to regional exposure measures, as was done in Finland,[17] without the problems of large sample size and individual case identification which would arise in regional case-control studies. The ecological study carried out in China[18] and described below is also an example of this approach.

One of the aims of ecological studies is to display the variation of disease rates and exposure across various populations in order to generate new hypotheses. For example, when international data are compared, the relationship between per capita fat and animal protein consumption, breast cancer incidence[19] and colorectal cancer[20] was found to be strong. However, because there is the possibility that these aggregate variables might not be representive of those for individuals,[3] it is unclear from these ecological studies whether fat consumption is actually involved in the aetiology of breast or colon cancer. Other factors such as socio-economic status also show strong correlations between both breast and colon cancer. However, wide geographical variations in the incidence of a number of diseases do indicate that one or more aetiological agents must be involved and that the determinants of these diseases are able to be manipulated.[21] If potential confounders (e.g. age, socio-economic status) are taken into account beforehand and controlled for, then some of these limitations can be overcome.

An example from China

China was found to have a large degree of geographical variation in mortality due to a number of causes.[22] Sixty-five counties were chosen because of their wide variation in mortality from 7 major cancers.[18] The selection of these counties also provided a large degree of variation for other causes of death. Within each of the 65 counties, 2 villages were randomly selected. Within each village, 50 adults (25 men and 25 women) were selected randomly in order to measure the within-county variability. Each person provided a blood and urine sample, food samples and information on their dietary, sociodemographic and lifestyle habits. This study was thus able to correlate mortality from 78 causes of death (measured in the mortality survey) with about 300 dietary, sociodemographic and lifestyle variables.[18]

Besides the obvious correlations, e.g. between the prevalence of cigarette smoking and lung cancer, a number of other significant correlations have emerged, for example between the prevalence of *Helicobacter pylori* IgG antibodies with gastric cancer mortality.[23] Fig. 1 shows the complexity of the relationship between *H. pylori* and gastric cancer. For example, the population in county 'U' has an almost 100% prevalence of *H. pylori* but a very low gastric cancer rate. By contrast, counties 'K', 'C' and 'V' have high seroprevalence rates of *H. pylori* and high gastric cancer mortality rates. What additional factors might be involved in gastric carcinogenesis? An answer might lie in comparison of, for example, dietary habits of populations in county 'U' with those of populations in counties with high gastric cancer rates ('K', 'C' or 'V'). The inconsistencies of correlation studies can also lead to more focused studies that attempt to interpret interactions between a number of exposures.

Other associations have included the relationship between hepatitis B antibody prevalence and cholesterol (but not aflatoxin level) and liver cancer,[24] strongly contested in ensuing correspondence,[25,26] schistosomiasis and colorectal cancer, the protective effect of green vegetable consumption on stomach and oesophageal cancer[18,23] and many others. This study is currently being repeated with a larger sample size for each county ($N = 100$).

Discussion

We are not advocating estimation of correlations in ecological studies as the best approach to the measure of associations. It may be preferable, for example, to fit an appropriate regression model and assess the magnitude of its coefficients.[2,27] One must also bear in mind the problem of multiple comparisons, a common method of analysis of nationally aggregated statistical data.[17,28] If one estimates a large number of associations and sets a 'significance level' of 5% then one would expect 5% of the associations which are actually zero, to be significant, just by chance. It would be important to specify the *a priori* hypotheses before interpreting any of the associations estimated in such studies. Incidentally, we have to be just as wary of 'data dredging' by the use of multiple regression variable selection techniques used increasingly in the analysis of data from observational studies.[29] An additional method of validation would be to test whether an ecological relationship also holds within population subgroups. Strong and consistent correlations were, for example, found between the prevalence of schistosomiasis and colorectal cancers between Chinese provinces and counties within each province.[30]

It is important, however, not to overinterpret the aetiological significance of data from ecological studies. The study design by Chen *et al.*[18] should therefore not be construed as a panacea for China's epidemiological problems. In fact, Chen and colleagues make no such claims and reiterate the need to couple these data with well-planned observational studies of individuals.

In China there was a concentrated effort to measure mortality rates over a 3-year period throughout the country.[22] Another advantage in China is that the rural population studied was relatively stable, and wide variations in mortality due to the diseases in question were found. For example there was a 25-fold geographical variation in gastric cancer rates[23] and a 45-fold variation in liver cancer rates.[26] In contrast only three regions, comprising a population equal to 0,25% of the sub-Saharan population, had collected any reliable mortality data for more than 1 year between the period 1986 and 1989.[31] In South Africa, up to 50% of deaths among blacks are not registered and, of the rest, over 20% are recorded as ill-defined, according to the Medical Research Council workshop on quality of mortality data (Tygerberg, 1990). Therefore the geographical distribution of deaths by cause might be, partly, a function of the efficiency of cause-of-death registration. A similar argument can be made for laboratory-diagnosed notifiable conditions. These might also reflect the distribution of access to laboratory facilities rather than the true underlying morbidity rate. While vital registration procedures may eventually improve, there is a need to measure the current burden of ill-health in South Africa and its geographical distribution.

A number of diseases are clinically readily identifiable and would bring most people to a medical doctor. These conditions would therefore be amenable to registration, and include most cancers and a number of communicable and non-communicable diseases.[32] The methodology for cancer registration, for instance, is now well established.[33] Despite difficulties in the identification of coronary events, the WHO-MONICA project was set up to investigate variations in incidence of cardiovascular disease in 39 collaborative centres.[34]

It may therefore be possible to set up a number of population-based disease registries (starting perhaps with cancer) and then add standardised procedures to record other identifiable conditions in areas with known denominators. What we are envisioning here is that if there is wide variation in incidence rates, ecological studies could be carried out that make use of information from such registries and combine it with exposure information which could be collected from sampling within each of the regions covered by each registry. If (and only if) a large geographical variation is found between different regions (both in disease prevalence and exposure), it may be possible to undertake a mini-Chinese study in South Africa to explore locally important disease-exposure associations.

An advantage of such a geographical design is that health planners tend to think geographically rather than on individual lines. Planners need to know which areas have high rates of particular disorders in order to intervene in a rational manner.

Conclusion

This discussion was not intended to undermine the invaluable contribution of prospective, case-control and cross-sectional studies in epidemiological research, and it is worth reiterating that by themselves ecological studies offer weak evidence of causal relationships.[35,36] The main aim of ecological studies is to show the spatial distribution of a disorder and to point to further observational studies of individuals. Our aim was to draw attention to the fact that in countries with limited financial resources, valuable insights may emerge from ecological studies and that they may overcome some of the problems of other types of observational studies.

We would like to thank Professor Richard Peto, Dr Calum Muir, Professor Syd Shapiro and Dr Chen Zhengming for their valuable comments on earlier drafts of this paper.

The National Cancer Registry is funded by the South African Institute for Medical Research, the Cancer Association of South Africa, the Department of Health and Population Development and the South African Medical Research Council.

APPENDIX

We illustrate below that regression on regional averages asymptotically adjusts for the regression dilution bias induced by measurement error.

Let Y_{ij} be the response measure for the j'th individual in the i'th region and X_{ij} the exposure measure for that same individual, where

$$X_{ij} = \mu + a_i + \eta_{ij}$$

and μ is the overall mean exposure level (across all regions), a_i is a random variable with mean 0 and variance σ_a^2 which reflects the between-region variability in exposure and η_{ij} is a random variable (independent of a_i) which reflects the within-region (between individuals) variation to exposure and has mean 0 and variance σ_η^2. We assume that

$$Y_{ij} = \beta\, X_{ij} + \epsilon_{ij}$$

where ϵ_{ij} are assumed to be independent of X_{ij} and to have mean 0 and variance σ^2.

To introduce measurement error into the model, assume now that one observes not X_{ij}, but x_{ij}, where

$$x_{ij} = X_{ij} + \delta_{ij}$$

i.e. x_{ij} is X_{ij} observed with error, where δ_{ij} (the random error) is independent of X_{ij} and has mean 0 and variance σ_δ^2. If one now fits the model

$$Y_{ij} = \beta'\, x_{ij} + \epsilon'_{ij}$$

then it is easily seen that

$$\text{covariance } (x_{ij}, Y_{ij}) = \beta' (\sigma_a^2 + \sigma_\eta^2 + \sigma_\delta^2).$$

It can, however, also be seen that

$$\text{cov } (x_{ij}, Y_{ij}) = \text{cov } (X_{ij}, Y_{ij}) = \beta\, (\sigma_a^2 + \sigma_\eta^2).$$
$$(\sigma_a^2 + \sigma_\eta^2 + \sigma_\delta^2).$$

Hence $\beta' = \beta\, (\sigma_a^2 + \sigma_\eta^2)/(\sigma_a^2 + \sigma_\eta^2 + \sigma_\delta^2) < \beta$ and, the larger the measurement error, the more the regression on the observer x_{ij} (as opposed to the 'true' X_{ij}) will be 'diluted'.

Now consider instead fitting a regression on regional averages:

$$\bar{Y}_{i.} = \beta''\, \bar{x}_{i.} + \epsilon''_i$$

where $Y_{i.} = \sum_j Y_{ij}/J$ = average response measure in the i'th region, similarly for $\bar{x}_{i.}$.

Then

$$\text{cov } (\bar{x}_{i.}, \bar{Y}_{i.}) = \beta'' (\sigma_a^2 + (\sigma_\eta^2 + \sigma_\delta^2)/J),$$

but

$$\text{cov } (\bar{x}_{i.}, Y_{i.}) = \text{cov } (\bar{X}_{i.}, \bar{Y}_{i.}) = \beta\, (\sigma_a^2 + \sigma_\eta^2/J)$$

and hence $\beta'' = (\sigma_a^2 + \sigma_\eta^2/J)/(\sigma_a^2 + (\sigma_\eta^2 + \sigma_\delta^2)/J)$.

It is easily seen, then, that

$$\beta' \leqslant \beta'' \leqslant \beta$$

and that $\beta'' \to \beta$ for large J, i.e. in words, the effect of regression dilution is reduced by the taking of regional averages and the regression coefficient in this case tends, for large within-region samples, to that without measurement error.

This argument was developed for the case of continuous X and Y variables, with X subject to measurement error. The case of dichotomous Y (presence/absence) of disease would involve modelling $P(Y = 1|X)$ by, say, a logistic model in X. Measurement error also leads to dilution of the coefficient in the logistic model.[37] The same line of argument as that used above can be followed approximately in this case to show that regional averaging also reduces regression dilution for the logistic model.

It must be noted that these arguments apply to a single exposure variable. When several exposure variables are modelled simultaneously, the effect of measurement error is more complex.[38,39]

REFERENCES

1. Robinson WS. Ecological correlations and the behaviour of individuals. *Am Sociol Review* 1950; **15:** 351-357.
2. Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health* 1982; **72:** 1336-1344.
3. Piantadosi S, Byar DP, Green SB. The ecological fallacy. *Am J Epidemiol* 1988; **127:** 893-904.
4. Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. *Int J Epidemiol* 1989; **18:** 269-274.
5. Willett W. *Nutritional Epidemiology*. Oxford: Oxford University Press, 1990.
6. Gardner MJ, Heady JA. Some effects of within person variability in epidemiological studies. *J Chron Dis* 1973; **26:** 781-795.
7. Fletcher C, Peto R, Tinker C, Speizer F. *The Natural History of Chronic Bronchitis and Emphysema*. Oxford: Oxford University Press, 1976.
8. McMahon S, Peto R, Cutler J, *et al.* Blood pressure, stroke and coronary heart disease: Part I. Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990; **331:** 765-774.
9. Irwig L, Glasziou P, Wilson A, Macaskill P. Estimating an individual's true cholesterol level and response to intervention. *JAMA* 1991; **266:** 1678-1685.
10. Rosner B, Spiegelman D, Willet WC. Correlation of logistic regression, relative risk estimate and confidence intervals for measurement error. The case of multiple covariates measured with error. *Am J Epidemiol* 1990; **132:** 734-745.
11. Rose G. Environmental health: problems and prospects. *J R Soc Phys (Lond)* 1991; **25:** 48-52.
12. Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984; **13:** 356-365.
13. Howe GR, Hirohata T, Hislop G, *et al.* Dietary factors and risk of breast cancer: combined analysis of 12 case-control studies. *J Natl Cancer Inst* 1990; **82:** 561-569.
14. Prentice RL, Sheppard L. Validity of international, time trend and migrant studies of dietary factors and disease risk. *Prev Med* 1989; **18:** 167-179.
15. Muir CS. Epidemiology, basic science and the prevention of cancer. Implications for the future. *Cancer Res* 1990; **50:** 6641-6648.
16. Ferris BG, Dockery DW, Ware JH, Speizer FE, Spiro R. The six-city study. Examples of problems in the analysis of data. *Environ Health Perspec* 1983; **52:** 115-123.
17. Teppo L, Pukkala E, Hakama M, Hakulinen T, Herva A, Saxen E. Way of life and cancer incidence in Finland. A municipality based ecological analysis. *Scand J Soc Med* 1980; suppl 19: 5-84.
18. Chen J, Campbell TC, Li J, Peto R. *Diet, Lifestyle and Mortality in China. A study of the characteristics of 65 Chinese counties*. Oxford: Oxford University Press, 1990.
19. Carroll KK. Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Res* 1975; **35:** 3374-3383.
20. Armstrong BK, Doll R. Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *Int J Cancer* 1975; **15:** 617-631.
21. Doll R, Peto R. *The Causes of Cancer*. Oxford: Oxford University Press, 1981.
22. Li J, Liu B, Li G, Chen Z, Sun X, Ring S. Atlas of mortality in the People's Republic of China. An aid for cancer control and research. *Int J Epidemiol* 1981; **10:** 127-133.
23. Forman D, Sitas F, Newell DG, *et al.* Geographic association of *Helicobacter pylori* antibody prevalence and gastric cancer mortality in rural China. *Int J Cancer* 1990; **46:** 608-611.
24. Campbell TC, Chen J, Liu C, Li J, Parpia B. Non-association of aflatoxin with primary liver cancer in a cross-sectional ecological survey in the People's Republic of China. *Cancer Res* 1990; **50:** 6882-6893.
25. Wild CP, Montesano R. Non-association of aflatoxin with primary liver cancer in a cross-sectional ecological survey in the People's Republic of China (Letter). *Cancer Res* 1991; **51:** 3825.
26. Campbell TC, Chen J, Liu C, Li J. Non-association of aflatoxin with primary liver cancer in a cross-sectional ecological survey in the People's Republic of China (Letter). *Cancer Res* 1991; **51:** 3826-3827.
27. Breslow NE, Enstrom JE. Geographic correlations between cancer mortality rates and alcohol — tobacco consumption in the United States. *J Natl Cancer Inst* 1974; **53:** 631-639.
28. Cohen BL. Ecological versus case-control studies for testing a linear no threshold dose-response relationship. *Int J Epidemiol* 1990; **19:** 680-684.
29. Greenland S. Modelling and variable selection in epidemiologic analysis. *Am J Public Health* 1989; **79:** 340-349.
30. Liu B, Rong Z, Sun X, Wu Y, Gao R. Study of geographical correlation between colorectal cancer and schistosomiasis in China. *Acta Academiae Medicinae Sinicae* 1983; **5:** 173-177.
31. Ruzica LT, Lopez AD. The use of cause of death statistics for health situation assessment: national and international experiences. *World Health Stat Q* 1990; **43:** 249-252.
32. Doll R, ed. *Methods of Geographical Pathology*. Oxford: Blackwell Scientific Publications, 1959.
33. Jensen OM, Parkin DM, Maclennan R, Muir CS, Skeet RG. *Cancer Registration. Principles and Methods*. World Health Organisation International Agency for Research on Cancer publication No. 95. Lyon: IARC, 1991.
34. World Health Organisation-MONICA project. Assessing CHD mortality and morbidity. *Int J Epidemiol* 1989; **18:** suppl 1, s38-s45.
35. Hill AB. The environment and disease: association or causation. *Proc R Soc Med* 1965; **58:** 1217-1219.
36. Susser M. *Epidemiology, Health and Society. Selected Papers*. Oxford: Oxford University Press, 1987.
37. Michalek JE, Tripathi RC. The effect of errors in diagnosis and measurement on the estimation of the ability of an event. *JASA* 1980; **75:** 713-721.
38. Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990; **1:** 421-429.
39. Kupper LL. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol* 1984; **120:** 643-648.