5

Between Epidemiology and Basic Genetic Research – Systems Epidemiology

Eiliv Lund
University of Tromsø
Norway

1. Introduction

Systems epidemiology can be considered as an attempt to implement functional genomic analyses into the common prospective design. Functional genomics cover research on genes, genomes and the products of genes such as gene transcripts (mRNA and microRNA) and proteins. Methods include gathering, integrating, and analyzing complex data from high throughput technologies such as genomics, epigenomics, transcriptomics, proteomics, and metabolomics (often collectively named "the 'omics"). A main goal is to build models to better understand the complex interactions taking place within cells, tissues or whole organisms, using mathematical, statistical and computational approaches. Some of these high throughput techniques can be run in available material, some need new biological sampling. The expansion of the information available through these methods has created a challenge for the analyses both in terms of laboratory analyses, statistical analyses and functional interpretations. At the same time it will mirror the current dichotomy in research between epidemiology and basic research. The goal of this chapter is to point to the alternative research direction of functional genomics created by the new technological opportunities. The time should have come for including functional measures in both blood and tissues in prospective observations studies of humans. With this as a background the design and current analytical approaches of the Norwegian Women and Cancer postgenome cohort will be discussed in relation to carcinogenesis. The chapter will deal more with study design and related aspects than with statistics or biology.

2. Background

Modern technology has over the last decade given epidemiologists the opportunity for expanding their field of science from the studies of associations between exposures and disease till gene-environment analyses of single nucleotide polymorphisms, SNPs, as part of molecular epidemiology (1). The genome wide association studies, GWAS, created both the need for huge collaborative efforts, high throughput technologies and novel statistical approaches due to the large number of comparisons done. One example of the collaborative efforts could be the Consortium of cohorts (2), and the adjustment of p-values to keep an adequate false discovery rate is an example of novel statistical methods for the GWAS analyses (3). As part of the gene-environment exploration the scientific approach has changed from single gene analyses based on biological knowledge of the function till

inductive or hypothesis generating approaches by looking at all genes simultaneously (1). This is also named the agnostic approach (4). So far the GWAS studies in cancer have discovered around 200 SNPs that most have a relative risk less than two. The post-GWAS strategy is under discussion, and the direction recommended is towards studies of functional aspects of these SNPs (4).

This development should be held against the common view behind the agnostic GWAS strategy which strongly points to the lack of exact biological information for making exact single genes or pathways approaches fruitful. In fact, the lack of *in vivo* derived information on most genes and pathways as part of carcinogenesis could hamper the search for mechanisms of carcinogenesis.

Another approach to expand the field of traditional epidemiology is systems epidemiology (5). This scientific discipline is the equivalent of systems biology, but performed in an epidemiological scale. In systems biology, high throughput 'omics' technologies are combined with computational analyses to investigate the metabolism of cells, tissues or organisms during health and disease. The aim of systems epidemiology is to study molecular mechanisms of disease in epidemiological studies. Systems epidemiology implicate better collections of biological material for functional studies and a carcinogenic model more relevant for epidemiology – an exposure driven functional model (6).

3. Status of functional genomics in epidemiology

The extent to which systems epidemiology is a realistic approach depends on the underlying assumption that blood and tissues communicate through gene expression and that the communication from cells undergoing a disease process through the blood might be trace signals from distorted metabolic pathways. This approach depends on adequately collected and stored biological material suitable for high throughput technologies used for studies of functional changes during the development of chronic diseases. Transcriptomics consists of two major classes of gene expression functions. mRNA is a copy or a messenger of the gene code information stored as DNA for the production of proteins in the cell. It is rapidly degraded by the Rnase. microRNAs are not coding for proteins, but regulates the expression of mRNAs. It is more resistance to degradation and can be used as biomarkers (7). New studies of the transport and delivery of microRNA are rapidly growing, strongly supporting the view that blood is an important channel for communication between cells. Thus, the basic assumption of systems epidemiology gains momentum.

The extent to which mRNA and microRNA are transported in the blood stream as information carriers can only be verified in humans through a prospective study design. There exist many studies with repeated blood samples with DNA from plasma/serum, but few with biological material suitable for gene expression analyses of mRNA simultaneously of peripheral blood and tumour tissue. One serious objection is the time frame for the function of gene expression which could differ, so the snapshot through one blood sample could give a confusing picture. But, those important cell regulations that are disturbed in the disease process should be expected to have some constancy over time since most exert the effects as a consequence of substantial exposures over a prolonged time. The success of this approach could depend on repeated measurements in order to be able to study changes in gene functions over a lifetime.

There is growing evidence that gene expression in peripheral blood reflects different lifestyle factors. Several cross-sectional studies of gene expression have been published highlighting numerous and interconnected pathways or gene sets affected in blood by defined lifestyle factors or exposure variables e.g. smoking (8), hormones (9) or organic pollutants (10). Important objections are the level of technical noise (11). Although blood gene expression profiling promises molecular-level insight into disease mechanisms, there remains a lack of baseline data describing the nature and extent of variability in blood gene expression in the general population. Characterizations of this variation and the underlying factors that most influence gene expression amongst healthy individuals play an important role in the feasibility, design and analysis of future blood-based studies. The number of studies with lifestyle exposures related to microRNA is absent and only a few studies exist for epigenetics "e.g". DNA methylation (12). In addition, a few case-control studies (13) have been published. So far these cross-sectional studies have not been transformed into prospective studies. At the same time a large number of studies have been publish based on clinical cohorts relating gene expression patterns in tumour tissue to survival and prognosis. Several studies have shown the usefulness of more functional classification of breast cancer (14). This might be important for etiological research as a means to improve the classification of breast cancer tumours.

4. An example of a biological model and the relationship to epidemiology: The two-stage model of carcinogenesis

In cancer epidemiology, the estimations of the carcinogenic multistage model is more than fifty years old (15). The situation today is not different from the early papers, namely that there is a lack of observational data of the stages of carcinogenesis. Due to this lack of observed data the parameters in the mathematical model can not be solved uniquely (16). At the same time the importance of fixing one parameter in the mathematical model has been stressed, this could be the duration or changes related to the last stage. There exist at least five models (15), some of them clearly more explored than others like the two-stage clonal growth model (17), figure 1.

The biological model considers that the carcinogenic process starts with a mutation which is a change in one of the DNA sequences of a gene. The cell with this mutation then will undergo a rapid growth named the clonal phase. A second mutation will be necessary in order to have a transformed cancer cell that will grow as a tumour through the promotional, last stage. Dependent on the exposure which drives the carcinogenic process a stop or withdrawal of exposure could bring the promotional stage into arrest or the cancer cells could die through a process named apoptosis.

5. The functional genomics of prospective studies – The globolomic design

The structure of a globolomic study design could be as shown in figure 2. On the left different sources of exposure information are given, from questionnaires, blood samples, tissue samples and pathological paraffin blocks. The differences between the traditional cohort study and the globolomic one is given at the right side of the figure 2. The richness of biological material multiplies the possible analytical strategies, at the same time the complexity is far beyond current epidemiological methodology.

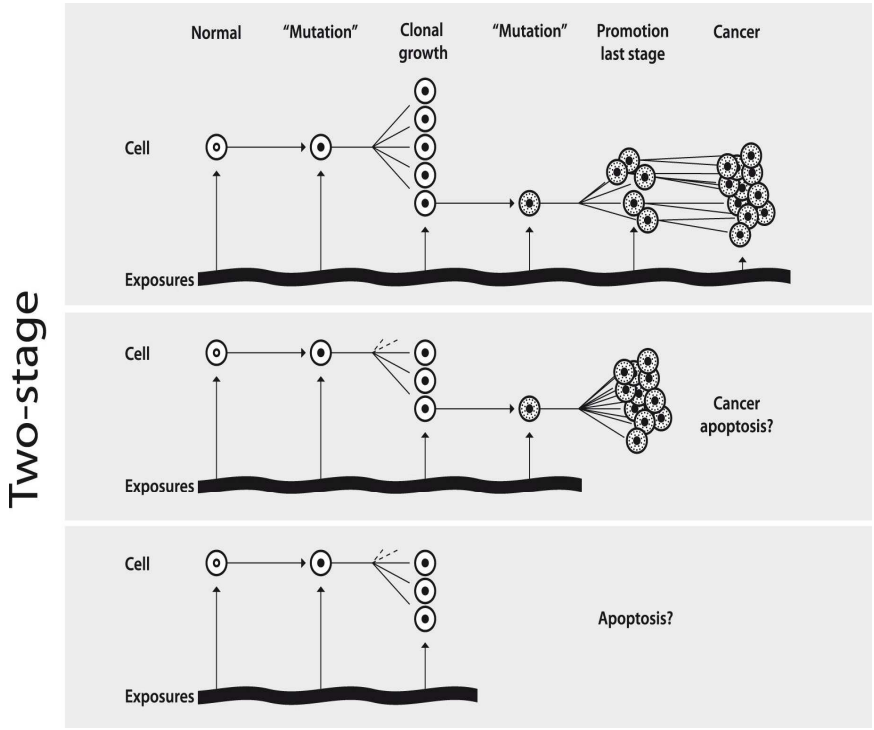


Fig. 1. Schematic description of the relationship between the clonal two-stage model and different scenarios of exposures.

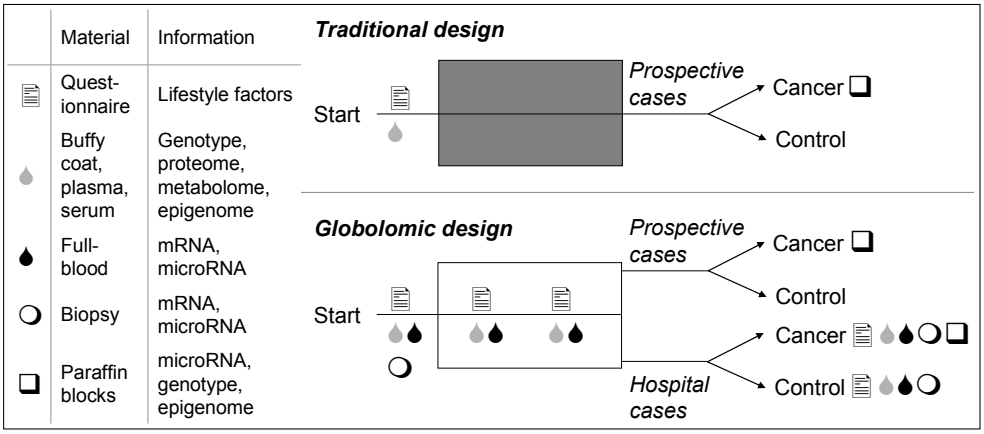


Fig. 2. The globolomic prospective study design.

As an example of the need for new and extended collection of both biological material and questionnaire information is given the structure of the Norwegian Women and Cancer cohort in Box 1, for more detailed information see (18).

1. Women sampled at random from the Norwegian population register, 172 000 women were enrolled.
2. Mailed letter of invitation and a questionnaire.
3. The postgenome biobank. Women born 1943-57 were eligible since they were invited or would be invited to participate in the Norwegian national mammographic screening program covering women 50-69 years, altogether 148 000.
4. Women were asked if they were willing to donate a blood sample to the study and at the same time give consent to update information on place of living. Approximately 95% of those returning the eight pages questionnaire answered yes to both questions.
5. Blood sampling. A package containing equipment for blood sampling and a two-paged questionnaire were mailed by groups of 500 random women. Participants brought the blood collection kit to their physicians office for blood sampling; one standard citrate tube and one collection tube containing a buffer for preservation of mRNA and microRNA. The blood samples give us access to mRNA and microRNA for gene expression, DNA from "buffy coat" for SNPs analyses, plasma for studies of metabolomics and proteomics.
6. The *passive follow-up* was completed through linkage to the national cancer registry in Norway based on the unique birth number and to registers of death and emigration. The information on cancer can then be used as end-points.
7. Collection of tumour biopsies through an *active follow-up* of all 148 000 women born 1943-57 at the time of diagnosis in collaboration with 10 of the major hospitals throughout Norway covered around 40% of the study sample. When a woman presented at the hospital with a lump in her breast, or one was diagnosed at the mammographic screening unit, all women were asked if they had participated in the NOWAC study. If they answered yes they were asked to give informed consent for a second biopsy for research use. At the same time they gave a blood sample and filled in a one page questionnaire.
8. For each of these women five random controls were drawn from the NOWAC postgenome cohort matched by age and date of the original blood donation. From these blood samples the same information can be extracted as from those blood samples collected originally. From the biopsy one can obtain microRNA, mRNA and tumour DNA.
9. For all cases of breast cancer the paraffin blocks stored in the pathological bio-banks are searched for to obtain microRNA and DNA.
10. Collection of breast tissue biopsies from healthy women participating in the NOWAC study and living in the north of Norway.

Box 1. Design, approaches and content of the Norwegian Women and Cancer postgenome study.

6. Challenges of systems epidemiology

The following will discuss several challenges raised by the introduction of functional genomics in prospective studies as part of systems epidemiology. The total amount of information is challenging the ordinary epidemiological analytical systems. From the DNA one can extract information for 500 000 SNPs, the same from tumour DNA, for each of the blood samples there will be a unique set of 25 000 gene expressions of mRNA and around 1 000 microRNAs. The methylation chips for epigenetic analyses cover around 500 000 variants. The number of measurements of metabolomics could be tens till hundreds. The proteomic screening analyses is just underway. Lastly, the questionnaire information could cover around 1 000 variables. In addition, there are scanned pictures from the microarrays and many files with technical information. Altogether, the storing and use of such large data sets will be dependent on computer science and cluster computers.

6.1 The nature of gene expression as exposure variable

In a prospective study gene expression could be classified as exposure. In a traditional design one would then use a Cox proportional model to estimate the relative hazard. This has typically been the procedure with the GWAS studies of SNPs. A SNP is a lifelong lasting characteristic that does not change over time or during follow-up. As such, SNPs could look as an ideal exposure variable being reliable and constant over the follow-up time. The analyses of gene expression in a follow-up study will be complicated by the possibilities of different population distributions throughout the follow-up period of the differences between cases and controls for each of the single of the 25 000 genes. Suppose we expect the gene expression in the controls to be similar over time. The hypothesis for a mutagen could be that the gene expression in the cases changes during follow-up as a consequence of events related to the disease process. We would then search for a change in the distribution of the differences in gene expression between cases and controls. These differences could have many potential distributions. The proportional hazard function would not be adequate.

The novel design gives us several challenges:

Challenge one: Biologically, the gene expression measured as the difference between the cases and the controls could either be the consequence of the exposure i.e. smoking changes the expression of a large number of genes, or the ongoing carcinogenic process due to the same exposures. This raises some methodological and statistical problems; how to estimate the changes in gene expression due to the carcinogenic process independently of the changes due to the carcinogen not necessarily linked to the carcinogenic process. If a mutation took place then the exponential increase in cancer cells could give a similar increase in expression of the affected genes. The differences in gene expression would then be an exponential function over time. Just putting both the gene expression variables and the exposure variables into the same model could give an unmeasured *over adjustment* of some of these variables. This can be handled by stratification which on the other hand would decrease the statistical power. Again, the analyses should be run agnostic before the information from basic research on gene function should be used.

Challenge two: As mentioned the traditional GWAS studies have been based mainly on the Cox proportional hazard model or using logistic regression analyses. The use of proportional hazard has the assumptions of proportionality and multiplicative risk estimation. There is no basic or epidemiological evidence of proportional hazard over time for the gene expression. In contrast, several other time-dependant models could be used.

The null hypothesis of no differences over time between the gene expression of cases and controls would in a linear model be closely parallel lines, eventually with the same beta-coeffisient.

One plausible model could be an increasing level of gene expression in cases compared to the controls as an effect of the mutations giving a clonal growth. This would give an exponential curve in an additive model or a straight line in a logarithmic model. There are many potential models that should be explored, but at the moment no strong preferences for the models exists from observational studies.

Challenge three: In traditional epidemiology including the gene-environment analyses of GWAS the search has been for the highest relative risks or the lowest p-values. This simple assumption does not hold for functional analyses. There is no evidence that important functional changes due to the disease process should be more clearly expressed than other ongoing cellular functions or effects of lifestyle. The search would be for genes that exerts a given time dependant pattern. The first analyses in systems epidemiology would be to sort the time-dependant functions of the 25 000 genes keeping in mind the consequences of the multiple testing. To sort out the highest p-values could remove very important information.

Challenge four: The time-dependant analysis of the gene expression could need new statistical tests and adaption of new functions for the follow-up studies. A major task both of design, laboratory work and statistical methods would be to improve the sensitivity of the analyses.

Challenge five: There is an obvious concern about the complexity of the total data structure of the functional information possible to obtain for a small number of cases and controls. This is a work that is ongoing in systems biology and several methods should be possible to adapt in order to improve the biological explanations of findings in the epidemiological studies.

7. Discussion

The design of a prospective study including trancriptomic options has only recently been implemented in the globolomic design of NOWAC. In the discussion of *pro et cons* for building new cohort studies the option for gene expression analyses are mostly neglected (19, 20, 21), but has been proposed by some (22).

The notion of gene-expression as exposures confronts the epidemiologists with approximately 25 000 possibly time-dependant exposure variables. This adds to the well known uniqueness of each woman's lifestyle. Consider data from the NOWAC study taking the mode of six well known factors that either increase or decrease the risk of breast cancer. Based on information from 172 000 women combining the mode value for age at menarche, parity, age at first full term pregnancy, age at menopause, age at first use of hormonal

replacement therapy and age at first use of oral contraceptive left no one to share the same lifetime exposure pattern. Even with so few variables the risk profiles of the women are highly different. It is under such conditions that the time dependant changes exert effects through the functioning of the genes. In order to focus on the overall importance of the exposures we would sum up over a person's lifetime the continuously changing lifestyle with both risk and preventive factors. The diversity of exposures gives a diversity of functional changes and an individual may at the same time have several potential carcinogenic processes ongoing even within the same tissue "e.g" the effect of smoking and radiation exposure on lung tissue.

It is well known that different exposures have different effect on the diseases. In cancer the carcinogenic process is different for exposures like radiation, chemicals, bacteria, and virus. In addition, several chemicals act as hormone imitators. There is no strong reason to believe in exact the same model of carcinogenesis for all exposures. Radiation hits the DNA in a different manner from use of hormones in postmenopausal women. Heterogeneity of exposures drives heterogeneity of functional changes and in the end the expression profiles.

7.1 Trans-etiological research

So far etiological or causal research has been done almost independently in basic cell biology and epidemiology except for the gene-environment analysis. One could call this a dichotomy, see Box 2.

	Epidemiology	basic genetic research
Common approaches		
Gene-enviroment analyses	exposure and gene interactions	
Bioinformatics	gene functions	
Dualism		
Model of carcinogenesis	multistage	mutational
Driving forces	exposures	mutations
Exposures	yes	mostly none
Methods	observational	experimental
Mechanistic/functional	no	yes, main focus
Scientific approach	whole genome scan	
Time relationship	prospective	cross-sectional end-point related
Causality	criteria for statistical association time order	experimental verification mRNA, oncogenes etc
Time relationship		

Box 2. Examples of common approaches and the dualism between epidemiology and basic cancer research.

In almost every aspect of scientific work these two disciplines have different views, methods and models. The expansion of functional genomics into epidemiology could improve the communications far beyond the current. While methodologies and designs of studies differ greatly, this could be considered as a natural consequence of the research fields. But behind this is the deeper conflict in science between those used to put up deductive hypothesis and test them in experiments versus the agnostic approach to the observational studies in epidemiology. The history of genomic analysis, SNPs, going from single gene studies over annotated genes till pathways analyses and ending up in GWAS clearly demonstrates the very different approaches scientifically in basic genetic research and epidemiology – from deductive designs of experiments till observational studies searching for associations. In order to improve collaboration between basic genetic research and epidemiology mutual understanding of methodological approaches in each discipline would be important.

8. Concluding remarks

A unique opportunity to expand design and interpretations of statistical associations in epidemiological studies has been given due to new technologies. At the same time this opportunity will depend on a closer collaboration between basic researchers in different biological disciplines and epidemiologist, giving us a possibility of a new trans-etiological research.

9. References

[1] Spitz MR, Bondy ML. The evolving discipline of molecular epidemiology of cancer. Carcinogenesis 2010; 31; 127-34

[2] Haiman CA, Dossus L, Setiawan VW, Stram DO, Dunning AM, Thomas G, Thun MJ, Albanes D, Altshuler D, Ardanaz E, Boeing H, Buring J, Burtt N, Calle EE, Chanock S, Clavel-Chapelon F, Colditz GA, Cox DG, Feigelson HS, Hankinson SE, Hayes RB, Henderson BE, Hirschhorn JN, Hoover R, Hunter DJ, Kaaks R, Kolonel LN, Le Marchand L, Lenner P, Lund E, Panico S, Peeters PH, Pike MC, Riboli E, Tjonneland A, Travis R, Trichopoulos D, Wacholder S, Ziegler RG. Genetic variation at the CYP19A1 locus predicting circulating oestrogen levels but not breast cancer risk in postmenopausal women. Cancer Res 2007; 67: 1-5.

[3] Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19: 368-75.

[4] Freedman ML et al. principles for the post-GWAS functional characterization of cancer risk loci. Nature Genetics 2011; 43; 513-18.

[5] Lund E, Dumeaux V: Towards a more functional concept of causality in cancer research. Int J Mol Epi Genet 2010; 1:124-133.

[6] Lund E. An exposure driven functional model of carcinogenesis. Med Hypoteses 2011 May 5. [Epub ahead of print]

[7] Cortez MA et al. MicroRNAs in body fluids – the mix of hormones and biomarkers. Nat Rev Clin Onc 2011; 8; 467-77.

[8] Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, Friend SH, Potter JD. Signature of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol Biomarkers Prev 2004; 13: 445-53.

[9] Waaseth M, Olsen KS, Rylander C, Lund E, Dumeaux V. Sex hormones and gene expression signatures in peripheral blood from postmenopausal women - the NOWAC postgenome study.BMC Med Genomics. 2011 Mar 31;4:29

[10] Terasaka S, Aita Y, Inoue A, Hayashi S, Nishigaki M, Aoyagi K, Sasaki H, Wada-Kiyama Y, Sakuma Y, Akaba S, Tanaka J, Sone H, Yonemoto J, Tanji M, Kiyama R. Using a customized DNA microarray for expression profiling of the estrogen-responsive genes to evaluate estrogen activity among natural estrogens and industrial chemicals. Environ Health Perspect 2004; 112: 773-81.

[11] Dumeaux V, Olsen SK, Paulssen RH, Børresen-Dale AL, Lund E. Deciphering blood gene expression variation – The postgenome NOWAC study. PLoS Genetics 2009 (in press)

[12] Vaissière T, Cuenin C, Paliwal A, Vineis P, Hoek G, Krzyzanowski M, Airoldi L, Dunning A, Garte S, Hainaut P, Malaveille C, Overvad K, Clavel-Chapelon F, Linseisen J, Boeing H, Trichopoulou A, Trichopoulos D, Kaladidi A, Palli D, Krogh V, Tumino R, Panico S, Bueno-De-Mesquita HB, Peeters PH, Kumle M, Gonzalez CA, Martinez C, Dorronsoro M, Barricarte A, Navarro C, Quiros JR, Berglund G, Janzon L, Jarvholm B, Day NE, Key TJ, Saracci R, Kaaks R, Riboli E, Hainaut P, Herceg Z. Quantitative analysis of DNA methylation after whole bisulfitome amplification of a minute amount of DNA from body fluids. Epigenetics. 2009 May 16;4(4):221-30. Epub 2009 May 24.

[13] Sharma P, Sahni NS, Tibshirani R, Skaane P, Urdal P, Berghagen H, Jensen M, Kristiansen L, Moen C, Sharma P, Zaka A, Arnes J, Sauer T, Akslen LA, Schlichting E, Børresen-Dale AL, Lönneborg A Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. Breast Cancer Res 2005; 7: R634-44.

[14] Sørli T. Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. Eur J Cancer 2004; 40: 2667-75

[15] Vineis P, Schatzkin A, Potter JD. Model of carcinogenesis: an overview. Carcinogenesis 2010; 31; 1703-9

[16] Cox LA, Huber WA. Symmetry, identifiability, and prediction uncertainties in multistage clonal expansion (MSCE) models of carcinogenesis. Risk Analysis 2007; 27; 1441-53.

[17] Moolgavkar SH, Day NE, Stevens RG. Two-stage model for carcinogenesis: epidemiology of breast cancer in females. JNCI 1980; 65: 559-569.

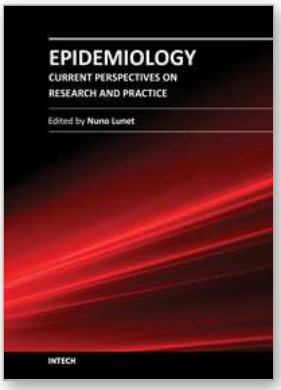
[18] Lund, E; Dumeaux, V; Braaten T; Hjartåker, Engeset D; Skeie, G; Kumle, M. Cohort profile: The Norwegian women and cancer study - NOWAC - Kvinner og kreft. International Journal of Epidemiology 2008; 37; 36-41

[19] Collins FS, Manolio TA. Necessary but not sufficient. Nature 2007; 445: 259.

[20] Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts: not worth the wait. Nature 2007: 445; 257-258.

[21] Colditz GA, Sellers TA, Trapido E. Epidemiology – identifying the causes and preventability of cancer? Nature Rev 2006; 75-83.

[22] Potter JD. Epidemiology informing clinical practice; from bills of mortality to population laboratories. Nat Clin Pract Oncol 2005; 2: 625-34.



Epidemiology - Current Perspectives on Research and Practice
Edited by Prof. Nuno Lunet

ISBN 978-953-51-0382-0
Hard cover, 208 pages
Publisher InTech
Published online 13, March, 2012
Published in print edition March, 2012

This special issue resulted from the invitation made to selected authors to contribute with an overview of a specific subject of their choice, and is based on a collection of papers chosen to exemplify some of the interests, uses and views of the epidemiology across different areas of research and practice. Rather than the comprehensiveness and coherence of a conventional textbook, readers will find a set of independent chapters, each of them of a great interest in their own specialized areas within epidemiology. Taken together, they illustrate the contrast between the attempt to extend the limits of applicability of epidemiological research, and the "regular" scientific activity in this field or an applied epidemiology. Epidemiologists with different levels of expertise and interests will be able to find informative and inspiring readings among the chapters of this book.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821