Commentary

Systems Epidemiology in Cancer

Eiliv Lund and Vanessa Dumeaux
Institute of Community Medicine, University of Tromsø, Tromsø, Norway

Abstract

Prospective studies in cancer epidemiology have served their study design over the last decades. In this context, current epidemiologic studies investigating gene-environment interactions are based on biobank for the analysis of genetic variation and biomarkers, using notified cancer as outcome. These studies result from the use of high-throughput technologies rather than from the development of novel design strategies. In this article, we propose the globolomic design to run integrated analyses of cancer risk covering the major -omics in blood and tumor tissue. We defined this design as an extension of the existing prospective design by collecting tissue and blood samples at time of diagnosis, including biological material suitable for transcriptome analysis. The globolomic design opens up for several new analytic strategies and, where gene expression profiles could be used to verify mechanistic information from experimental biology, adds a new dimension to causality in epidemiology. This could improve, for example, the interpretation of risk estimates related to single nucleotide polymorphisms in gene-environment studies by changing the criterion of biological plausibility from a subjective discussion of in vitro information to observational data of human in vivo gene expression. This ambitious design should consider the complexity of the multistage carcinogenic process, the latency time, and the changing lifestyle of the cohort members. This design could open the new research discipline of systems epidemiology, defined in this article as a counterpart to systems biology. Systems epidemiology with a focus on gene functions challenges the current concept of biobanking, which focuses mainly on DNA analyses. (Cancer Epidemiol Biomarkers Prev 2008;17(11):2954–7)

The description of the human genome in 2001 (1, 2) has created a paradigmatic change in cancer research. In cancer epidemiology, the vast efforts in gene-environment studies seemingly illustrate such a change with the use of high-throughput technologies that largely increase both the number of analyses in each individual and the potential number of persons involved in the analysis. The design of prospective cohort studies has remained unchanged for about 60 years (3). Thus, there might be great potential for innovative designs introducing new technologies and directly link epidemiology and basic biological research. Based on our experience (4, 5), we will propose a new research discipline, systems epidemiology, that seeks to integrate pathways analyses into observational study designs to improve the understanding of biological processes in the human organism. Systems epidemiology is the observational counterpart to systems biology, which has many definitions, such as “a discipline that seeks to determine how complex biological systems function by integrating experimentally derived information through mathematical and computing solutions” (6). One could eliminate both terms and use the common term “systems science” (7), but this would not emphasize the emerging approaches and designs necessary for the optimal use of new technologies. Here, we will introduce the globolomic design offering new analytic opportunities to study systems epidemiology, and describe possible consequences for the interpretation of causality.

Current Genetic Epidemiologic Studies, Systems Epidemiology, and the Globolomic Design

The novel study design inherent to systems epidemiology corresponds to the expansion of gene-environment epidemiologic studies with analyses of the transcriptome. In this setting, gene expression profiles could provide us with a biological mechanistic insight, and/or be used as biomarkers, of exposure and outcome. In the current prospective study design, questionnaire information about exposure, biomarkers, and single nucleotide polymorphisms are all considered as variables that are analyzed in association with registered cancer as outcome. One example could be The European Prospective Investigation into Nutrition and Cancer (8). The prospective design is mainly a consequence of the development of epidemiologic methods in the 1980s and 1990s, with an increasing focus on selection and recall bias in case-control studies (9). The cohort design reduces these biases. It has been pointed out that many of the studies of gene-environment interactions did not have an adequate design (10, 11). Observational studies can and do produce findings that either spuriously enhance or downgrade estimates of causal associations between exposures and disease (12).
Under the current design of genetic epidemiologic studies, the transcriptome as a biomarker of exposure at the initial stage of the study, and/or of outcome at the end of follow-up, remains a “black box.” While the implementation of transcriptome analysis will give a more complex epidemiologic design, it also gives a more complete biological model system. Up-to-date technologies have provided us with unique opportunities for interactive studies of the genetic predisposition (DNA-genome), the expression of genes (transcriptome), proteins (proteome), and metabolites (metabolome). Information on genes, gene variants, gene expression and modification, proteins, and signaling and metabolic pathways can be integrated across many levels in addition to lifestyle information collected by questionnaires.

Systems epidemiology could be seen as a discipline that merges epidemiologic research with biological mechanistic analysis by investigating gene expression patterns related to metabolic pathways. Systems epidemiology creates the need for novel designs, and our proposal is a globolomic study design. The concept of the globolomic design is close to the “population laboratory” described in 2006 as a potential future design by Potter (10). Pathway analyses in the globolomic design of observational studies could link epidemiology and systems biology more closely. Its success depends upon the ability to look beyond a single biological context and the unidimensional view of gene expression (13) within an epidemiologic study setting. With the globolomic design one can look at the observed gene expression profiles resulting from the complex real life situation, with hundreds of different exposures interacting with genetic predisposition and the risk of cancer. The novelty of this design is the possibility to run expression analyses of mRNA and microRNA (miRNA) in peripheral blood and tumor samples. This will open up the “black boxes” of the gene expression on the exposure and outcome side, which can also be used as biomarkers of exposure and outcome.

Thus far only rudiments of knowledge exist about transcriptomics in cancer epidemiology, and the pursuit of complex models and interactions has been hampered by small sample sizes and inadequate study design in many molecular applications (11, 14). A few small experimental studies of gene expression analyses related to exposure have identified gene expression patterns in human peripheral blood related to benzene (15), metal fumes (16), dioxin (17), and smoking (18) exposure. In a recent cross-sectional analysis of expression profiles in blood and adipose tissue, a marked correlation was found in adipose tissues, but less so for blood (19). A recent study investigating gene expression profiles related to hormone therapy use, identified as known human carcinogens by the International Agency for Research on Cancer in 2005 (20), revealed very few genes in blood with an increase or decrease in expression greater than 1.5-fold (21). The sensitivity needed to detect low-level gene expression changes in whole blood (22) is further complicated by different sources of technical variability (work in progress).

The recent discovery of miRNAs has seemingly complicated pathway analysis based on gene expression profiles, but studies of miRNA functions in the regulation of the main cellular processes involved in carcinogenesis open up a novel research area (23). The prognostic performance of miRNA profiles in cancer has been found to be at least as good as that of mRNA profiles (24). Only one large-scale prospective study has thus far included biological material for transcriptome analysis (5). Similarly, no epidemiologic studies, not even small cross-sectional analyses, have collected adequate material to conduct miRNA analyses.

New Design Opportunities and Challenges

The implementation of gene expression analysis in a prospective study increases the analytic options in a remarkable way (Fig. 1). First, information from a biobank with DNA, RNA, plasma, and questionnaires collected at time of inclusion with follow-up through notification of cancer diagnosis gives the opportunity to analyze blood samples before diagnosis in a prospective design. Subsequently, the collection of biological material from the tumor tissue and blood of cancer patients is a challenge because this has to be done at time of diagnosis and before any treatment. Depending on the source of the cohort members, different strategies might apply. With the identification of cohort cases one can approach matched controls from the cohort. By combining the information from these two time points the gene expression profiles and biomarkers for the same person can be compared over time. Further, the tumor tissue can be used for comparing gene expression profiles in cases and controls and for etiologic research of the taxonomy based on gene expression profiles. This design also offers the

![Figure 1](image-url). The globolomic study design.
opportunity for verifying the gene expression of genes associated with increased or decreased risk of cancer found in gene-environment studies, either in blood or tumor tissue. Due to the current understanding of cancer as a multistage process (25) with a long latency time, the interpretation of expression in blood at any time should take into account the different stages of the carcinogenic process. Ideally, the expression analysis should cover a long time span for each participant with the possibilities of looking at gene expression changes over time, implying repeated collection of blood samples.

The globolomic design will produce an enormous amount of data. Microarrays for genome-wide genotyping use up to 1 million single nucleotide polymorphisms for the genetic characterization. The questionnaire information could cover hundreds of variables. The possible number of analyses of the metabolome may be in the hundreds. The mRNA analyses cover more than 25,000 genes and around 500 miRNA species (26). Finally, the same amount of information from the genome, metabolome, and transcriptome is obtained at time of diagnosis in blood samples and from the transcriptome in tumor tissue. The most compelling challenge will be to integrate data, spanning multiple levels of the biological scale, and environment information from the questionnaires to create systems epidemiology.

Causality in Systems Epidemiology

In epidemiology, causality is mostly discussed through the use of certain criteria of causality, originally developed by Hill (27). Some of these criteria are linked to the design of the studies and can be judged based on quantified information (temporality, strength of the association, dose-response, and specificity of exposures or outcomes). Other criteria are externally related (consistency with other studies, prediction, relation to health statistics, lack of alternative explanations, and analogy). Another set of criteria for judging causality in gene-environment studies, including large-scale genotyping, has been introduced, namely, the amount of evidence, the extent of replication, and protection from bias (28). The last, and in this context, important criterion described by Hill is biological plausibility or coherence with biological knowledge. This criterion has thus far been discussed in terms of possible mechanistic aspects based on information taken from basic biological research. The mechanistic information is based on in vitro or animal experiments mainly with a “reductionistic” design, a rather simplistic view today (29). The need for a human verification of basic mechanistic research in relation to gene-environment studies could be illustrated by the classification of carcinogens by the International Agency for Research on Cancer (30). For any chemical or lifestyle factor to be classified as a human carcinogen, epidemiologic evidence is almost mandatory. Of the products classified as human carcinogens by the International Agency for Research on Cancer, only three have been placed in this class solely on the basis of mechanistic studies and with no epidemiologic evidence, dioxin being the most known of these (31). The obvious reason for this precaution is the lack of generalization from in vitro or animal experiments. Due to the complexity in human lifestyle and genetic disposition, strong human evidence must be assured before any public health or clinical implications can be concluded from mechanistic studies. Causality of cancer in basic biological research is less debated, but a recent review considers oncogenes, tumor suppressor genes, and miRNA as causes of cancer (32). Thus, these two different concepts of causality might merge in the globolomic design by the potential for a quantitative estimation of pathway changes involved in the above-mentioned genes by analyzing the expression of mRNA and miRNA in blood and tumor tissue in an observational study setting.

The potential of systems epidemiology can be seen in light of the upcoming major research field of gene-environment studies. The results of these analyses strongly focus on pathways analysis and mechanistic reasoning. Genetic epidemiology has been tested over the past decade with frequent failures to find robust replicable associations between genetic variants and common diseases (33). An example of the relationship between biological complexity and statistical power in prospective studies is the analysis of 5,356 invasive breast cancer cases and more than 7,000 controls in the National Cancer Institute Breast and Prostate Cancer Cohort (34). The study found an association between genetic variation at the CYP19A1 locus with a 10% to 20% increase in endogenous estrogen levels, but not with breast cancer risk. This could be due to the lack of statistical power to find an increased breast cancer risk of 1.07 to 1.15. The authors concluded with a quantitative statement and uncertainty about the use of genotype information for clinical use. In a globolomic design the lack of consistency in the risk for breast cancer and endogenous levels of hormones could be highlighted by including analyses of mRNA and miRNA that might help us to elucidate the pathways of hormone action. If small changes in estrogen levels increase breast cancer risk, then one could hypothesize that this should be reflected in the gene expression.

The second example is taken from a large genome-wide association study on more than 4,500 cases of lung cancer and 7,000 controls (35). The conclusion was that there is a susceptibility locus nicotine acetylcholine receptor unit on chromosome 15q25. The implication for a specific disease mechanism is suggestive and points to possible effects on tumor growth through these receptors. This recent publication of the large-scale genome-wide association study is something of a relief for those who have pioneered the common variant–common disease hypothesis and has required a large-scale collaboration.

Tracking down the genetic variants that regulate metabolic pathways of relevance to common diseases will, hopefully, provide a better understanding of molecular pathophysiology through transcriptomic, metabolomic, and proteomic research. The likely existence of far more gene-environment interactions than we already have knowledge of provides justification for a combined genetic, genomic, and environmental epidemiologic approach to understanding causation.

These two examples of gene-environment analyses clearly show that a further increase in statistical power will not solve the complicated questions of causes of cancer, but reveal a need for more relevant design of epidemiologic studies in relation to the rapidly growing
knowledge and improving technology in basic research. The information from prospective studies are associations or risk estimates that form the basis for discussion on causality. The mechanistic information from in vitro experiments reveals something about pathways or mechanisms, and gives some indirect clues about causality. Currently, there is no way from one side to the other except by using assumptions. These two concepts of causality cannot merge until we have a common scientific design including both aspects. The globolomic design could be a meeting place where the risk estimation of gene-environment studies could be expanded, with pathway analysis or gene expression analysis adding the mechanistic understanding to the concept of causality in epidemiology.

Conclusion

New knowledge and technology have over the last few years given cancer epidemiology the possibility of a new research discipline, systems epidemiology, that could move epidemiology to the research front of technology and basic science, and by this route create a shift in the paradigm of cancer epidemiology. This expansion of cancer research could foster new study designs like our proposed globolomic design that is a linked prospective and nested case-control study with the possibility of analyses of all -omics, including the transcriptome in blood and tissue samples. As a potential consequence, the criterion of biological plausibility, which today is discussed using knowledge derived from basic research based on in vitro or animal experiments, could in an appropriate design be quantified as patterns of gene expression related to given pathways.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

6. Institute of Systems and Synthetic Biology. Imperial College London. www3.imperial.ac.uk/systemsbio.
Systems Epidemiology in Cancer
Eiliv Lund and Vanessa Dumeaux
Cancer Epidemiol Biomarkers Prev 2008;17:2954-2957.

Updated version
Access the most recent version of this article at:
http://cebp.aacrjournals.org/content/17/11/2954

Cited articles
This article cites 33 articles, 9 of which you can access for free at:
http://cebp.aacrjournals.org/content/17/11/2954.full#ref-list-1

Citing articles
This article has been cited by 6 HighWire-hosted articles. Access the articles at:
http://cebp.aacrjournals.org/content/17/11/2954.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.