EDITORIAL

Epidemiology in the Post-Genomic Era

K S Chia,* *FAMS, MD, FFOM (Lond)*

To the clinical specialists, epidemiology appears to be a hyphenated word associated with their clinical specialty: for example, cancer-epidemiology, cardiovascular-epidemiology, infectious disease-epidemiology and occupational-epidemiology. It is a toolbox to unravel the distribution and determinants of diseases within their specialty in the community. The tools are predominantly quantitative and the underlying principle is to compare distribution patterns between cases and non-cases or the exposed and non-exposed. The assumption is that if an exposure has a biological relationship with the health outcome, its distribution will differ between those who have and do not have the outcome. Intuitively, this makes sense but unfortunately, the converse is not necessarily true. As a result, it had spawned great methodological developments in the areas of handling bias and confounding.

The focus on unravelling differences in distribution within observational studies riddled with bias and confounding, without paying much attention to the biological basis of the association, has contributed to the era of 'black box' and subsequently risk factor epidemiology. Such an approach is characterised by a search for multiple antecedent factors and linking them with the outcome without the necessity of intervening factors.[1] Those who advocate such a strategy sees it as a 'unique virtue' and has made significant contributions.[2] For example, the association between arsenic exposure and lung cancer was made for many years without an animal model for arsenic carcinogenicity.[3]

However, in the light of developments in molecular biology and the unravelling of the human genome, taking refuge within the 'black box' will only generate many more inconclusive studies of associations like coffee drinking and cancers or hip fractures.[4] Of course, the aim is not to replace a narrow approach with another. One must also be cautious and avoid 'throwing the baby out with the bathwater'.[5] In a passionate appeal, Susser argues that 'the focus of the (black box) paradigm is too narrow to cope with the future' and the 'failure to invent or recognise a new paradigm exacts the penalty of stagnation and inertia'.[1]

In the last two decades, there has been a new twist in the twining of words with epidemiology. Instead of the more comfortable traditional terms like cancer and infectious disease epidemiology (by disease group) or even occupational and nutritional epidemiology (by exposure group), newer terms like 'biochemical epidemiology' and 'molecular epidemiology' herald the availability of new laboratory techniques for epidemiological research. Genetic epidemiology is another term in vogue in the light of the human genome project.[6]

The newer laboratory techniques bring with them the promise of more accurate exposure and outcome assessment. Moderate non-differential misclassification of exposure and outcome degrades the observed rate ratios (relative risks), even in well-designed and well-conducted epidemiological studies. Measurement of DNA adducts in target tissues or surrogate tissues like white blood cells are believed to be more accurate estimates of the 'biologically effective dose'. Before an absorbed compound can form an adduct, it is usually metabolised by enzymes which may differ due to inherited polymorphic DNA sequence variation or because of gene upregulation or downregulation caused by other exogenous agents. Therefore, DNA adduct levels may not correlate well with the environmental level of carcinogen. It is a better indicator of the internal dose that may cause DNA mutation when the cell replicates. Unfortunately, there are few clear examples of adduct measurements shedding light on new exposure-disease association.[7]

Despite questions as to the validity of various biomarkers for population-based research,[8] one of the dreams in molecular epidemiology is to confirm 'signature' mutations caused by specific carcinogens. The specific guanine to thymidine (G to T) mutations of codon 249 in the p53 gene in hepatocellular carcinoma from Africa and China[9,10] brought great excitement and a boost for molecular epidemiology.

* *Associate Professor*
Department of Community, Occupational and Family Medicine
National University of Singapore
Address for Reprints: Dr Chia Kee Seng, Department of Community, Occupational and Family Medicine, Faculty of Medicine MD3, National University of Singapore, 16 Medical Drive, Singapore 117597.

Another example is the discovery of high-penetrance susceptibility genes such as BRCA1 and BRCA2.[11] It gave epidemiologists a more 'accurate' instrument to define inheritable susceptibility to diseases. Theoretically, analyses could now be stratified into high- and low-risk groups and the associations between 'environmental' exposure and disease in the low-risk group will be magnified. In reality, for most common diseases, the prevalence of such high-penetrance genes is small. The inclusion of those with the high-penetrance genes will have only a modest effect on the overall rate ratios.

Of greater challenge is the low-penetrance genes and the so-called 'complex' diseases. With the availability of the DNA sequence of all human genes and high-throughput genotyping and DNA chips, variants of one or many candidate genes can be compared among cases and controls. The statistical debate on 'multiple comparisons' and 'multi-collineraity' will have to be revisited. It may no longer be reasonable to arbitrarily partition the *P* value or to use the conventional argument of including only 'biologically plausible' candidate genes for data analyses. Rethinking of traditional study designs[12] to cope with the availability of multiple, repeated and related biomarkers may be necessary.

The availability of biological markers of genetic and environmental factors has opened the door to explore complex gene-gene and environment-environment interactions as well. Large sample sizes, in the order of 1000 or more cases, are needed if traditional markers are used. This could be reduced if relevant biomarkers are used.[7]

Such studies of gene-environment interactions should become central in the post-genomic era. There have been intense debates among epidemiologists on the definition and source of interaction as well as the methods for detecting them.[13-18] Conceptually, gene-environment interaction can be defined as a different effect of an environmental factor on the disease outcome for persons with different genotype. From a statistical point of view, the difference in effect can be measured on an additive (rate difference) or multiplicative (rate ratio) model. Rothman[13] argues that the additive model is the best model for interaction as it measures interaction between two factors that are part of different causal mechanisms. Further, if the aim is to predict disease load in a population, the additive model is more suitable. The multiplicative model is more appropriate if the primary goal is to unravel disease aetiology (as is the case of most gene-environment interaction studies) and is suitable for factors that act through the same mechanism or the same stage of a multistage process.[15]

Ottman[19] described five biologically plausible models of gene-environment interaction. The genotype results in an increase in expression of an environmental factor, which can also be produced non-genetically. A modification to this model is where the genotype must be present for the environmental factor to have an effect. Conversely, the third scenario is where the environmental factor has no effect with low risk genotype. The fourth model is when both environmental and genetic factors are needed together to increase risk and finally both factors can independently increase risk but modifies each other's effect when present. Further, she attempted to integrate these biological models with epidemiological models. The integration is far from ideal and highlights the current inadequacy of both the biological and epidemiological models.

Just as the 'black box' era had spurred the development of methodological issues to handle bias and confounding, the post-genomic era will probably stimulate further insights in our understanding of interaction (or effect modification). For example, we have several local epidemiological studies exploring the interaction between polymorphism of metabolic enzymes and environmental carcinogens in lung cancer among non-smokers. A paper has been published[20] with two others in preparation. To cope with analysing different levels and mode of interactions, more complex mathematical models like multilevel (hierarchical) regression[21] may be necessary. The principle of parsimony in regression models may have to be questioned if the 'truth' is more complex.

Epidemiology has evolved through the decades. It survived the death of the Miasma Theory (where diseases are attributable to foul air) and transformed itself into infectious disease epidemiology with the increasingly popular Germ Theory. When chronic diseases came to the forefront, epidemiology took on the 'black box' and 'risk factor' approach. Now in the era of gene sequencing and molecular techniques, it has put on the new names of genetic epidemiology and molecular epidemiology. In this transition, it is going through a soul-searching process.[22] Although it sees the need to move 'upstream' and incorporate molecular and mechanistic understanding, it will need to maintain its original population perspective.[5] It will ultimately realign itself and contribute the necessary concepts and methodology to the understanding of health and diseases at the population level in this new molecular and genomic revolution. In time to come, perhaps the term molecular or genetic epidemiology will become superfluous. The day when molecular techniques and genomic information are thoroughly incorporated into epidemiological studies is near. Epidemiology would have then evolved into the study of the distribution and *interaction of genetic and environmental factors* of diseases with the ultimate aim of controlling them in the most cost-effective means.

REFERENCES

1. Susser M. Does risk factor epidemiology put epidemiology at risk? Peering into the future. J Epidemiol Community Health 1998; 52:608-11.
2. Savitz D A. In defense of black box epidemiology. Epidemiology 1994; 5:550-2.
3. Landrigan P J. Arsenic. In: Rom W N, editor. Environmental and Occupational Medicine. Baltimore: Williams & Wilkins, 1983:473-9.
4. Feinstein A R, Horwitz R I, Spitzer W O, Battista R N. Coffee and pancreatic cancer: the problem of etiologic science and epidemiological case-control research. JAMA 1981; 246:957-61.
5. Pearce N. Epidemiology as a population science. Int J Epidemiol 1999; 28:S1015-8.
6. Khoury M J, Beaty T H, Cohen B H. Fundamentals of genetic epidemiology. New York: Oxford Univ Press, 1993.
7. Ross R K, Yuan J M, Yu M C, Wogan G N, Qian G S, Tu J T, et al. Urinary aflatoxin biomarkers and the risk of hepatocellular carcinoma. Lancet 1992; 339:943-6.
8. Pearce N, Sanjose S, Boffetta P, Kogevinas M, Saracci R, Savitz D. Limitations of biomarkers of exposure in cancer epidemiology. Epidemiology 1995; 6:190-4.
9. Hsu I C, Metcalf R A, Sun T, Welsh J A, Wang N J, Harris C C. Mutational hotspot in the p53 gene in human hepatocellular carcinomas. Nature 1992; 350:427-8.
10. Bressac B, Kew M, Wands J, Ozturk M. Selective G to T mutations of p53 gene in hepatocellular carcinoma from southern Africa. Nature 1991; 350: 429-21.
11. Brody L C, Biesecker B B. Breast cancer susceptibility genes. BRCA1 and BRCA2. Medicine 1998; 77:208-26.
12. Miettinen O S. Etiologic research: needed revisions of concepts and principles. Scand J Work Environ Health 1999; 25:484-90.
13. Greenland S, Rothman K J. Concepts of interaction. In: Rothman K J, Greenland S, editors. Modern Epidemiology. Philadelphia: Lippincot Raven, 1998:329-42.
14. Walter S D, Holford T R. Additive, multiplicative and other models for disease risks. Am J Epidemiol 1978; 108:341-6.
15. Siemiatycki J, Thomas D C. Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 1981; 10:383-7.
16. Miettinen O S. Causal and preventive interdependence: elementary principles. Scand J Work Environ Health 1982; 8:159-68.
17. Greenland S, Poole C. Invariants and noninvariants in the concept of interdependent effects. Scand J Work Environ Health 1988; 14:125-9.
18. Pearce N. Analytical implications of epidemiological concepts of interaction. Int J Epidemiol 1989; 18:976-80.
19. Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol 1990; 7:177-85.
20. Seow A, Zhao B, Poh W T, Teh M, Eng P, Wang Y T, et al. NAT2 slow acetylator genotype is associated with increased risk of lung cancer among non-smoking Chinese women in Singapore. Carcinogenesis 1999; 20:1877-81.
21. Greenland S. Principles of multilevel modelling. Int J Epidemiol 2000; 29:158-67.
22. Saracci R. Epidemiology in progress: thoughts, tensions and targets. Int J Epidemiol 1999; 28:S997-9.