



## **COMMITTEE ON CARCINOGENICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT**

### **STATEMENT ON THE INVESTIGATION OF INTERACTION BETWEEN GENOTYPE AND CHEMICALS IN THE ENVIRONMENT ON THE INDUCTION OF CANCER**

#### **Background to review**

1. The Committee was asked by the Department of Health to review the available information on the interaction between genotype and exposure to chemicals in the environment and the induction of cancer. The Committee was asked to provide advice on the methods of epidemiological research used in this area, the approaches to identifying genes of interest for such studies and the evaluation and significance of these data for cancer risk assessment.

2. The Committee was aware of the major technological advances in rapid DNA sequencing which had been published by the Human Genome Project (HGP) (<http://www.ornl.gov/hgmis/project/.html>) and the Environmental Genome Project (EGP) (<http://www.niehs.nih.gov/envgenom/>). [See Introduction to review and glossary for explanation of abbreviations and terms used in this statement] A draft scaffold sequence for the human genome was published in February 2001 and it has been proposed by HGP that a complete high quality DNA reference sequence will be available by 2003. These projects have as their major goal, the diagnosis, prediction and intervention in diseases where there is a genetic contribution to the cause of disease. However the EGP is focused on the role of genes implicated in cellular responses to environmental chemicals. Ambitious projects have been set up by EGP, for example, to identify Single Nucleotide Polymorphisms (SNPs) for up to 30,000 genes, new statistical methods to aid in the evaluation of the interaction effects of carcinogen metabolism and bioinformatics tools to assist in the evaluation of the large amounts of data generated from epidemiological studies. These, and other developments<sup>1-3</sup> are likely to lead to a rapid increase in the published information on the interaction between genotype and exposure to chemicals in the induction of specific cancers. It was considered timely to examine the questions raised by the Department of Health in order to draw conclusions on what advice could be given.

3. The Committee considered that it was necessary to set out a discussion of the key terms in the text of the statement. The Committee agreed that a concise “non-technical” summary was also required which should provide a glossary of key terms. Members considered it appropriate to discuss the critical areas of the review, particularly the design of epidemiology studies for genotype-environment interactions for specific cancers, the identification of genes of interest and risk assessment, before providing advice and suggestions for further research. The Committee was provided with a set of detailed papers drafted by the DH Toxicology Unit at Imperial College

of Science, Technology and Medicine for use in their discussions.<sup>4-7</sup> The relevant papers will be published on the COC website ([www.doh.gov.uk/coc.htm](http://www.doh.gov.uk/coc.htm)) .

## **Introduction to review**

### *Background to terminology\**

*\*(see HGP and EGP internet sites (para 2 for addresses) and refs 4-7)*

4. The human genome comprises all the genetic material (i.e sequence of DNA) in the 23 pairs of chromosomes present in all somatic nucleated cells in the body. Within the genome, the gene is the fundamental physical and functional unit of heredity. A gene is an ordered sequence of DNA located in a particular position on a particular chromosome that encodes for a specific functional product(s) (i.e. a sequence of RNA which may be translated to give a protein(s) which, with any subsequent necessary posttranslational modification, gives the functional protein). One key development arising from the expansion in DNA resequencing work described in paragraph 2 of this statement has been the recognition that the human genome does vary considerably between individuals (i.e it is subject to considerable interindividual variability). Thus it is estimated that approximately 1 in every 300-500 base pairs will differ between any two individuals. Variation in the DNA sequence of a particular gene between individuals comprising a single nucleotide difference is called a Single Nucleotide Polymorphism (SNP). The variations in the genome between individual ranging from SNPs, differences in small sequences of DNA, up to whole chromosomes are collectively referred to as “genotypic variation”. During its discussions the Committee was principally concerned with SNPs. These may have no impact on the function of the encoded gene products and are called “non-functional but in some cases SNPs do result in variation between individuals in the function, e.g. qualitative and/or quantitative changes in protein function. This is referred to as phenotypic variation”. The phenotype of an individual is defined as the observable physical biochemical or physiological characteristics of that individual.

5. Within the genome, SNPs can be found in the coding region of a gene; i.e. functional DNA, (cSNPs), in potential regulatory sequences, i.e. peri-genic regions (pSNPs) or in intervening stretches of DNA with no apparent function (intergenic DNA; iSNPs). The term genetic polymorphism is often used to indicate phenotypic variation and as such is frequently used in association with genetically-determined variations in the metabolising capacity for chemicals. The changes in DNA sequence responsible for metabolic polymorphisms are often SNPs. It is now easier and more pragmatic to identify the genotype in large numbers of individuals by DNA sequencing rather than to elucidate phenotype (i.e. measure the expression and function of genes) and thus it is possible that fewer studies of phenotype will be undertaken. However it is the phenotypic expression of genes that is most likely to be important with regard to the interaction between a gene and an environmental chemical in the induction of specific cancers.

6. The development in DNA sequencing techniques has allowed for the rapid and easy identification of SNPs, and hence closer examination of whether there is an interaction between the occurrence of a particular SNP in an individual and chemical exposure that is associated with adverse health effects such as cancer. Many

epidemiological studies have investigated associations between cancer incidence and polymorphisms of the enzymes responsible for the metabolism of chemical carcinogens since many carcinogens require metabolic activation. It is therefore logical to suggest that variation in metabolism of these chemicals will accord with changes in risk of cancer development. The Committee reviewed several examples, e.g. *N*-acetyltransferase 2 (NAT 2) and exposure to tobacco smoke associated with bladder cancer, and glutathione-S-transferase M1 (GSTM1) and exposure to tobacco smoke associated with lung cancer.<sup>7</sup> However the Committee felt it was important also to consider target genes other than those associated with the metabolism of chemicals. A discussion paper was therefore drafted on this topic.<sup>6</sup>

7. The term “penetrance” is used in this statement to describe the frequency with which carriers (e.g. of a particular genotype) develop cancer, i.e. the ratio of carriers who develop cancer compared to all carriers. Inherited cancer genes are considered to be “high penetrance” if affected individuals have a high probability of getting cancer. An example is the breast cancer susceptibility gene BRCA1, where the lifetime cumulative risk of cancer in individuals carrying specific mutations within this gene has been estimated as approximately 90%. However “high penetrant” genes are usually rare, i.e. their prevalence in the population is low. The genotypes under consideration in this statement are considered to be of low penetrance, i.e. the increase in risk of cancer is very low. However their prevalence in the population can be very high (e.g. 40-50% of the population as with *N*-acetyltransferase 2 slow acetylator (NAT2) allele polymorphism and GSTM1 null polymorphism). The Committee noted that the penetrance and prevalence of genotypes that were of importance to carcinogenesis induced by chemicals could vary and thus this should be considered in strategies to identify genes for research.

8. The use of the term “interaction” has been considered in detail.<sup>8-10</sup> There are two ways in which this term has been used in the scientific literature: either to describe a biological model of interaction between two or more factors in the aetiology of disease or to describe the statistical concept of interaction which describes the patterns of disease risks. Thus for genotype-environment interactions, a biologically significant effect infers that there is evidence for or there is a presumed (as yet unknown) biological consequence arising from the function of a particular gene variant and exposure to chemical(s) on the risk of cancer. The degree of statistical interaction can be measured in two ways, depending on whether it is the differences (i.e. additive scale) or ratios of risks (i.e. multiplicative scale) that are of interest. An illustrative numerical example is given in the Annex at Table 1 based on the lifetime risks of lung cancer. Further explanation of the example is given below in paragraphs 30-31 that concern risk assessment.

#### *Discussion of critical areas to be considered*

9. Members noted the rapid increase in the number of publications on genotype-environment interactions and in particular those concerning the potential impact of metabolic polymorphisms. They were also aware of the suggestion that genetic screening could be used to identify individuals carrying a particular genotype or to identify chemicals to which individuals should avoid exposure. The Committee

therefore agreed there were two questions which needed to be addressed during the review namely;

- i) The extent to which subgroups of the population can be identified, who because they have a particular genotype, are at greater risk of developing cancer, when exposed to particular chemicals.
- ii) Is it appropriate or desirable to use genetic screening to identify individuals with a particular genotype of importance to chemically induced cancers.

10. Members agreed that in order to consider these two questions, it would be important to review the epidemiological methods used and to comment on the significance and potential value of the results from these studies for risk assessment. An integral part of this consideration would be to provide advice on the numbers of individuals required in such studies. This review would also provide advice on gene selection, and the formulation of hypotheses for future epidemiological investigations

11. Members agreed that a further critical area for review involved the discussion of the nature of the interactions between genotype and exposure to chemicals resulting in an increased risk of cancer. The objective was to define criteria which could help to assess whether an interaction existed between a particular genotype and exposure to chemicals leading to an increase in the frequency of cancer that was significant for public health. This will assist in differentiating between genotype-environment interaction associated with increased risk of cancer, and those which are chance findings, and therefore not relevant to risk assessment.

12. It would then be important to define the data necessary to assess the potential impact of interactions between genotype and exposure to chemicals and if possible to estimate potential numbers of cancer cases that might be involved. It would also be important to provide advice on the prospects and desirability in regard to the suggestion for genetic screening. In this regard, the Committee was of the view that a number of critical genotype-environment interactions have yet to be discovered.

13. The Committee agreed that any conclusions should be prefaced with a discussion of the uncertainties in the assessment.

#### **The Assessment of Genotype-Environment interaction studies<sup>4,5</sup>**

14. The recognition that many cancer susceptibility genes are likely to be of low penetrance has led to the evolution of two major study designs for the assessment of gene-environment interactions.

- i) epidemiological studies of candidate susceptibility genes (gene characterisation studies) and,
- ii) genetic association studies (gene discovery studies).

15. In the first, the influence of known polymorphisms (or SNPs) on cancer risk is determined, usually in case-control or cohort studies, whilst in the second, cases and controls are genetically screened in an attempt to identify a clear difference in one or more gene loci. Most studies of the interaction between genotype and exposure to chemicals to date have involved the first design, but the increasing availability of dense SNP (single nucleotide polymorphism) maps and the technology to perform large numbers of genotyping tests is making the second design much more feasible.

#### *Gene characterisation*

16. Gene characterisation studies should involve the *a priori* selection of candidate genes to be included in the study protocol before the investigation is initiated. A number of different study designs had been used including case-control (with a variety of methods for choosing controls), cohort and case-only. Members noted that the use of case-only designs was relatively recent and could provide an estimate of the strength of interaction between genotype and exposure to chemicals but such studies assumed independence between the effects of genotype and exposure to chemicals. Members felt that overall many of the available genotype-environment interaction studies suffered from flaws in design and/or interpretation, reducing their potential value in cancer risk assessment. The Committee considered that apart from the limitations often found in epidemiological studies such as measurement error, bias and confounding, a key concern for many published studies was the absence of clearly stating the *a priori* hypotheses to be tested before undertaking the epidemiological investigation. The reliance of many research groups on *post hoc* analyses of sub-groups after data had been generated could yield biased statistical analysis of the multiple comparisons common in such studies. Members considered that the *a priori* hypotheses under investigation should be clearly stated in publications, perhaps even lodged with a third party before the analysis. The Committee reviewed study designs used for case-control gene characterisation studies and agreed that good study design would require careful selection of cases and controls from the same population, adequate exposure assessment, appropriate analysis strategy, and power calculations of necessary study size (given assumptions on penetrance, relative risk of disease and prevalence of susceptible genotype).

17. The Committee reviewed some model calculations for a case-control study design based on published approaches to the consideration of study size. The calculations assumed that genotype and chemical exposure had independent effects on cancer risk, there was no matching of cases and controls, and a multiplicative interaction was of interest, a baseline cancer rate of 0.001 and the odds ratio (OR) for cancer in non-susceptible subjects from exposure was 1.5.

18. These calculations provide evidence to show that many of the currently published case-control studies are of insufficient size to identify moderate interactions between genotype and exposure to chemicals in the induction of cancer.

Table 1. Number of subjects required in case-control studies.

Proportion of susceptible genotype in population	Strength of interaction to be detected	Number of subjects (equal number of cases and controls)
0.5 (50%)	2x	2215
	5x	485
0.2	2x	3891
	5x	1017
0.05	2x	13902
	5x	3949

Two-tailed test of null hypothesis,  $P < 0.05$ ; power, 0.8. (Calculations were performed using the "Power" program described by García-Closas and Lubin (1999), American Journal of Epidemiology, **131**, 552-566.

### *Gene Discovery*

19. Until recently, gene discovery designs have not been used widely in genotype-environment studies of cancer. This is because of the impracticalities involved in screening the very large numbers of subjects and alleles that would be necessary for the detection of genes (i.e. sequence variants such as SNPs) of low penetrance. However, as indicated above, rapid advances in both knowledge and technology are making such study designs more feasible, and several groups have commenced or are about to commence such studies.

20. The Committee considered that there were at least four broad categories of gene for which it was reasonable to hypothesise that genotype-environment interactions might be of importance with regard to cancer.<sup>6</sup>

Table 2 Categories of gene associated with Genotype/Environment interactions.

Category of gene	Examples
Increased metabolic activation and/or reduced detoxication, elimination	Cytochrome P450 isozymes (e.g CYP1A1 and CYP2E1). Glutathione S-transferases (e.g GSTM1 and GSTT1). N-acetyltransferases (e.g NAT1 and NAT2). P-glycoprotein transporters
Reduced capacity for DNA repair	Very few studies to date have examined DNA repair capacity, suggestions include base and nucleotide excision repair genes
Immune surveillance	No data available in respect of immune surveillance, suggestions include human leukocyte antigen complex (HLA).
Increased potential for cell proliferation and survival resulting from alterations in control of cell cycle and apoptosis.	No data available. Suggestions for cell cycle control include cyclinD1 and HRAS1 and for apoptosis Bcl-2.

21. Members were aware that there was a large number of publications which had reported investigations of genetic polymorphisms of enzymes of metabolic activation and detoxication (e.g. cytochrome P450 dependent monooxygenases, glutathione-S-transferases and *N*-acetyltransferases) and some studies had included investigations of the combined effect of two or more metabolic polymorphisms for these enzymes. Comparatively few studies had investigated variants of the other categories of genes identified by the Committee, and the extent to which these might interact with environmental chemicals was unknown.

22. The Committee considered that it was difficult to know how to prioritise the search for gene variants with increased risks for environmentally induced cancers as this could plausibly involve many thousands of such variants. However, members believed that the benefits of improving and developing technology could result in this exercise being practical and useful in the future. Members noted that the Environmental Genome Project had identified similar categories of genes for inclusion in the first phase of its project on gene discovery.

23. Members agreed that as understanding of the pathways and genes involved in the biological processes critical to cancer development increases, the number of candidate genes within those pathways that may be relevant to study for interaction with environment would increase rapidly.

*Criteria for assessing interactions between genotype and environment in the aetiology of cancer<sup>7</sup>*

24. The Bradford-Hill criteria for causality<sup>11</sup> have been used in the past to investigate single risk factors (environmental or genetic) by both this Committee and the WHO International Agency for Research on Cancer (IARC reviewed studies of selected metabolic polymorphisms and susceptibility to cancer. Whilst there was no formal attempt to establish causality, the conclusions reached were based on the Bradford-Hill criteria.<sup>12</sup>)

25. In contrast to investigations of single factors, the Committee agreed that, consideration of genotype-environment interactions referred to the assessment of whether the occurrence of a particular genotype and exposure to chemicals was associated with an increased frequency of cancer that was of significance for public health. The Committee agreed that an assessment of genotype-environment interactions should ideally require information on the gene variants(s) under consideration, the mechanism of carcinogenicity of the chemical under consideration and evidence to link all of this information together to form a reasoned case.

26. Members acknowledged, however, that it was likely that future investigations would examine the potential role of several hundreds or thousands of genes simultaneously and felt that, for the assessment of genotype-environment interaction studies, initial emphasis would be placed on the strength and consistency of the association. This would require demonstration of consistency in both gene discovery and characterisation studies and preferably by several different methods in adequately conducted gene characterisation studies. There would also need to be a plausible rationale for the mechanism of carcinogenicity for the chemical under consideration.

This assessment should ideally include information on phenotype, but it is recognised that such information may not always be available.

27. Members also highlighted the potential problem of random co-inheritance (i.e. linkage disequilibrium), where alleles of one gene (associated with increased risk) are inherited with specific alleles of adjacent genes (unrelated to risk) giving the false impression that these latter genes were also causally associated with increased risk. Therefore, in the absence of knowledge of which genes are co-inherited, it would be important to have some understanding of the mechanism of carcinogenesis of an environmental chemical before any final conclusions could be reached.

28. Thus the Committee agreed that a tiered approach to the assessment of genotype-environment interactions was required as outlined in paras 24-27. It would only be possible to undertake a quantitative risk assessment if there was compelling evidence that a true interaction existed.

### **Risk Assessment**<sup>7</sup>

#### *Significance of genotype-environment interactions for public health*

29. The Committee agreed that a full assessment of the significance of genotype-environment interactions with regard to chemically induced cancer required considerable information to be available. Thus ideally data on the prevalence of chemical exposure, prevalence of susceptible genotype and the cancer incidence rate in those exposed with and without the genotype and in those non-exposed with and without the genotype of interest. However, some useful measures of the size of an interaction and its impact can be estimated if relative risks (or odds ratios) are available instead of incidence rates (See Annex). In many instances such data would not be available and thus any evaluation would be based on incomplete data..

30. The Committee reviewed a worked example where appropriate data were available, namely lung cancer, smoking and GSTM1 polymorphism. The rationale for choosing this example was that the particular cancer is common, there is good agreement regarding the exposed attributable fraction for lung cancer associated with smoking (cf 90%) and the polymorphism chosen was common (i.e. 50% of population).

31. The results of the model calculations are given in Tables 1 and 2 in the Annex at the end of this statement. Any measure of *population impact* needs to take into account the prevalence of both the high-risk genotype and the environmental exposure, as well as the risks of disease in each exposure combination. One approach would be to simply use this information to work out the numbers of cancer cases who would be predicted to occur in each exposure subgroup (as outline in para 29, exposed with and without the genotype and non-exposed with and without the genotype of interest) and hence the population impact. Another approach would be to use the *population attributable fraction* (PAF). For a single risk factor this is usually considered to be the fraction for exposure to a single factor of disease in a population that might be avoided if the exposure had not occurred (or by eliminating that exposure). The model calculation estimates the PAF for all potential exposure

subgroups. It is also possible to calculate the *exposed attributable fraction* of disease which provides information on the fraction in the exposed subgroups which might be avoided by eliminating exposure. The results shown in the Annex Table 1 suggest that for the example of GSTM1 polymorphism and lung cancer there is a slight benefit to the population impact (in reduction of numbers of individuals with lung cancer) in targeting smokers with GSTM1 but only if effective intervention is feasible. This has important implications when reviewing the practicalities of screening (see para 34 below).

32. Members were aware that to date most studies had investigated the interaction between metabolic polymorphism (i.e variation in the metabolising capacity for chemicals) and cancer.<sup>11</sup> The majority of studies (using either case-control or cohort methods) report modest increases in relative risk in exposed individuals with the susceptibility genotype. Without information on the factors outlined in paragraph 29 above, it would be difficult to derive conclusions on the significance for public health of the genotype-exposure interaction. The Committee concluded that the available data on metabolic polymorphisms had failed to demonstrate any consistent strong association between any one gene-environment interaction and cancer risk and therefore the interactions studied to date were likely to be of little importance for public health or risk assessment.<sup>11</sup> However, this did not exclude the possibility that genotype-environment interactions with a significant impact on cancer risk would be identified in the future.

33. The existence of an association between a genotype and chemical exposure in the induction of cancer (e.g. a phenotype that results in enhanced metabolic activation of the chemical) could provide supporting epidemiological evidence in the identification of human carcinogens. It would be important to demonstrate a plausible biological association between the mechanism of carcinogenesis and the genotype/phenotype measured.

#### *Significance for genetic screening*

34. A final measure of impact which can be derived is the Number Needed to Screen.<sup>13</sup> This combines together the prevalence of genotypes, the risks of cancer in each subgroup and the reduction in risks which could be achieved by screening identified individuals where effective intervention was possible. The Committee noted that the data for the example used in model calculations suggested that there was little value in screening for GSTM1 polymorphism. The Committee reaffirmed its view that when the environmental exposure is smoking, the only appropriate public health intervention was to aid all smokers in giving up smoking. The Committee reviewed some further published calculations which confirmed that it was impractical to screen for these low-penetrant genotypes in the general population.<sup>13</sup> The Committee noted that, at present, there is little value for risk assessment in screening for the genotypes identified to date in gene-environment interaction studies. This is because the number of individuals with the genotype of interest who would develop cancer would be small, whereas there would be large numbers of individuals with the genotype of interest who would not develop cancer. Members were also concerned that screening for such low penetrant genotypes was undesirable in that the information would not have any significant predictivity of individual risk of cancer. In addition, other risk factors for cancer such as diet and smoking were likely to be of

much greater importance in determining individual risk. The Committee also noted that there were considerable ethical, legal and social issues to be considered with regard to any proposal for screening which were beyond the scope of this review.<sup>14</sup>

### **Discussion and conclusions**

35. The Committee noted that there were considerable practical difficulties in assessing the significance for public health of the currently available genotype-environment interaction epidemiology studies of cancer. These related to the size and design of the investigations and the absence of clearly set out *a priori* hypotheses as an essential part of study design. Many of the studies published also had limited power to detect genotype-environment interactions. The Committee was aware of the rapid advances in DNA re-sequencing in the last few years which meant that many potential candidate genes and genotypes/SNPs for investigation in genotype-environment cancer studies were being identified and many more would be forthcoming. The Committee noted that several projects had been set up under the U.S Environmental Genome Project and other initiatives to address these issues.

36. The Committee agreed that there was a need to assess all the available information and to consider if there was compelling evidence that a true genotype-environment interaction existed before using the information in quantitative risk assessment. The Committee agreed that an assessment regarding a genotype-environment interaction should ideally require information on the gene variants(s) under consideration, the mechanism of carcinogenicity of the chemical under consideration and evidence to link all of this information together to form a reasoned case. It was necessary to consider the possibility of linkage disequilibrium. With regard to the assessment of genotype-environment interactions, initial weight should be placed on the strength and consistency of the association. There was also a need to provide a reasoned case linking the mechanism of carcinogenicity of the chemical with the genotype under consideration. This would involve some knowledge of the function of the gene in question. Members agreed it would be valuable to have full information on the phenotype including characterisation of the function of the gene product and information on chemical-phenotype interaction but acknowledged this might be a lengthy process. An interim assessment could be drawn on basic information on gene function.

37. The Committee concluded that the available data on metabolic polymorphisms had failed to demonstrate any consistent strong association between any one gene-environment interaction and cancer risk and therefore the interactions studied to date were likely to be of little importance for public health or risk assessment. However, this did not exclude the possibility that genotype-environment interactions with a significant impact on cancer risk would be identified in the future. The Committee discussed the likely scenarios under which genotype-environment interactions might be of significance for public health and also commented on the feasibility and desirability for genetic screening for low penetrance gene variants in genotype-environment interactions. It was acknowledged that future gene discovery studies might identify genotype-environment interactions involving gene variants of significant penetrance and prevalence for cancer and thus the literature on this subject should be kept under review.

38. The Committee agreed the following overall conclusions.
- i) The most appropriate study designs for gene characterisation investigations will vary according to study purpose. Many of the currently available studies are either too limited in size or relied on *post hoc* analyses to highlight selected results. Ideally, studies should include information on phenotypic variation, but it is unlikely that such data would be available for all candidate genes selected for investigation. It is essential that such studies should involve *a priori* hypotheses. There is an argument that such hypotheses should be lodged with a third party before epidemiological investigations are undertaken.
  - ii) The rapid development of DNA sequencing techniques means that many gene discovery studies will become available in the future. There is currently no clear rationale for gene selection for gene discovery studies, other than to state broad categories of genes that could be prioritised for consideration (such as metabolic activation, DNA repair and immune surveillance, cell proliferation and cell cycle control).
  - iii) Before the results of genotype-environment interaction studies can be used in risk assessment (either for the identification of susceptible populations or identification of human carcinogens), it is necessary to establish whether there is a reasonable case to infer that the genotype-environment interaction is associated with a real and important increased frequency of cancer. A tiered approach has been recommended. Initially the strength and consistency of evidence from the epidemiological studies should be considered. In addition information to establish if there is a credible link between the mechanism of carcinogenicity for the chemical and the function of the gene and genotype under investigation should be considered. This assessment should ideally include information on phenotype, but it is recognised that such information may not always be available.
  - iv) The Committee concluded that the available data on metabolic polymorphisms had failed to demonstrate any consistent strong association between any one gene-environment interaction and cancer risk and therefore the interactions studied to date were likely to be of little importance for public health or risk assessment.
  - v) There is little value in using genetic screening to identify individuals with particular genotypes of interest for carcinogenesis induced by environmental chemicals.
  - vi) The possibility cannot be excluded that genotype-environment interactions involving gene variants of significant penetrance and prevalence might be identified through gene discovery investigations in the future or that combinations of genotypes might result in

significantly greater interaction with chemicals in the induction of cancer.

- vii The Committee recommended that it was important to keep this subject under review particularly in the light of expected developments arising from the Environmental Genome Project based in the U.S.A. and other initiatives in this area.

**June 2002**

## **References**

1. Taioli E (1999). International Collaborative Study on genetic susceptibility to environmental carcinogens. *Cancer Epidemiol Biomarkers Prev.* Aug;8(8):727-8.
2. Single Nucleotide Polymorphism Consortium TSC ltd. [www.snp.cshl.org/](http://www.snp.cshl.org/)
3. [www.ncc.nih.gov/plan2002/concerngb/scpagenes.htm](http://www.ncc.nih.gov/plan2002/concerngb/scpagenes.htm)
4. DH Toxicology Unit (2000). Genetic susceptibility to environmental carcinogens pp 1-10, 17 June 2000.
5. DH Toxicology Unit (2001). Criteria for the design of gene-environment epidemiology studies, 5 March 2001.
6. DH Toxicology Unit (2001). A review of potential target genes for susceptibility to carcinogenesis. 5 March 2001.
7. DH Toxicology Unit (2001). A review of how gene-environment studies should be used in risk assessment process. 5 March 2001.
8. Yang Q and Khoury MJ (1997). Evolving methods in Genetic Epidemiology. III. Gene-Environment Interaction in Epidemiologic Research. *Epidemiologic Reviews*, 19, (1), 33-43.
9. Ottman R (1996). Gene-Environment Interaction: Definitions and Study Designs. *Preventive Medicine*, 25, 764-770.
10. Clayton D and McKeigue PM (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *The Lancet*, 358, 1356-1360.
11. Bradford-Hill A (1965). The Environment and Disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295-300.
12. IARC (1999). Metabolic polymorphisms and susceptibility to cancer. Edited by Vineis P, Malats N, Lang M, d'Errico A, Caporaso N, Cuzick J and Boffetta P. IARC Scientific Publications No 148, Lyon, France.

13. Vineis P, Schulte P and McMichael AJ (2001). Misconceptions about the use of genetic tests in populations. *The Lancet*, 357, 709-712.
14. Coughlin SS et al (1999). BRAC1 and BRAC2 gene mutations and risk of breast cancer: public health perspectives. *American Journal of Preventive Medicine*, 16, 91-98.

## Annex

### Calculating and interpreting genotype-environment interaction: an example using lung cancer, smoking and GSTM1 polymorphisms

The statistical definition of a genotype-environment interaction is that the effect of genotype on disease risk varies with the level of exposure to an environmental factor, or vice versa. The degree of statistical interaction can be measured in two ways, depending on whether it is the *differences* or *ratios* of risks that are of interest. For simplicity, it is assumed that the variables measuring disease, exposure and genotype are all dichotomous. An illustrative numerical example is given in Table 1 based on the lifetime risks of lung cancer. The environmental factor is cigarette smoking (+ = Yes – = No), and the genotype of interest is the GSTM1 polymorphism (+ = null – = wild). The lifetime risk of lung cancer in non-smokers with the low-risk genotype was assumed to be 1.2%, and the lung cancer risks in the other subgroups were plausible estimates from the literature.

Table 1 Lung cancer risks in each subgroup

Environmental factor (Smoking)	Genetic factor (GSTM1)	Lifetime lung cancer risk (%)	Relative risk
+	+	$R_{++}$ 16.0	$RR_{++}$ 16/1.2 = 13.33
+	-	$R_{+-}$ 12.0	$RR_{+-}$ 12/1.2 = 10
-	+	$R_{-+}$ 1.6	$RR_{-+}$ 1.6/1.2 = 1.33
-	-	$R_{--}$ 1.2	Reference subgroup
Whole population		4.55	

#### Measures of genotype-environment interaction

If *differences* in lung cancer risk are of interest, the *Additive* measure of interaction contrasts the difference between the risks of those with the high and low risk genotype who are exposed to the environmental factor ( $R_{++} - R_{+-}$ ), to the same difference for those unexposed to the environmental factor ( $R_{-+} - R_{--}$ ).

i.e. there is no Additive interaction if  $(R_{++} - R_{+-}) = (R_{-+} - R_{--})$

or equivalently, in terms of relative risks, if  $(RR_{++} - RR_{+-}) = (RR_{-+} - 1)$

A measure of *Additive* interaction is therefore  $(RR_{++} - RR_{+-} - RR_{-+} + 1)$  and a value of 0 denotes no additive interaction<sup>1</sup>.

In Table 1  $(RR_{++} - RR_{+-}) - (RR_{-+} - 1) = (13.33 - 10 - 1.33 + 1) = 3.0$ , so there is some interaction on an additive scale, since there is a larger difference between the cancer risks for the null and wild genotypes for smokers than for non-smokers.

If the *ratios* of the risks are of interest, the *Multiplicative* measure of interaction contrasts the ratio of the risks between those with the high and low risk genotype who are exposed to the environmental factor ( $R_{++} \div R_{+-}$ ), to the same ratio for those unexposed to the environmental factor ( $R_{-+} \div R_{--}$ ).

i.e. there is no Multiplicative interaction if  $(R_{++} \div R_{+-}) = (R_{-+} \div R_{--})$

or equivalently in terms of relative risks  $RR_{++} = (RR_{+-} \times RR_{-+})$   
 A measure of *Multiplicative* interaction is therefore  $RR_{++} \div (RR_{+-} \times RR_{-+})$ , and a value of 1 denotes no multiplicative interaction<sup>1</sup>.  
 In Table 1  $RR_{++} \div (RR_{+-} \times RR_{-+}) = 13.33 \div (10 \times 1.33) = 1$ , so there is no interaction on a multiplicative scale, since the cancer risk increases by the same ratio between the null and wild genotypes in smokers and non-smokers.

NB If the data come from case-control studies, then the absolute disease risks will not be available. However, relative risks (RR) can be estimated by odds ratios (OR) and the measures of genotype-environment interaction above can be estimated using the appropriate OR.

### Population impact of genotype-environment interaction

These measures of genotype-environment interaction give an idea of the size and type of any interaction between two factors, but don't permit an assessment of the impact of the interaction on the whole population or selected subgroups. A variety of such measures are described below. Any measure of population impact needs to take into account the prevalence of both the high-risk genotype and the environmental exposure, as well as the risks of disease in each genotype-exposure combination. In this example, the prevalence of smoking was taken to be 25%, while that of the null genotype was 50%, and they were assumed to occur independently.

Table 2 Measures of population impact of genotype-environment interaction

Environmental factor (Smoking)	Genetic factor (GSTM1)	Exposure prevalence (%)	Cases in population of 50 million ('000s)	Population attributable fraction (%)	Exposed attributable fraction (%)
+	+	P <sub>++</sub> 12.5	1000	40.7	92.5
+	-	P <sub>+-</sub> 12.5	750	29.7	90.0
-	+	P <sub>-+</sub> 37.5	300	3.3	25.0
-	-	P <sub>--</sub> 37.5	225	-	-
Total population		100	2275		

One approach would be to assess the population impact simply by calculating the numbers of lung cancer cases that would be predicted to occur in each subgroup: multiplying together the absolute cancer risks in each genotype-exposure subgroup by the appropriate prevalence. Table 2 shows the predicted numbers of subjects in a population of 50 million, who would get lung cancer at some point in their lifetime, for each genotype-exposure subgroup. It can be seen that of the 2.275 million cases expected in the whole population, the highest numbers of cases occur in the GSTM1-null & smokers subgroup, followed by the GSTM1-wild & smokers subgroup. The lack of multiplicative interaction between GSTM1 and smoking has increased the risk of lung cancer for the GSTM1 -null genotype by 33.3% in both smokers and non-smokers. However, the addition interaction means that given the 10-fold extra risk for smokers, the GSTM1-null genotype has had a more noticeable impact on the expected number of lung cancer cases among smokers than non-smokers: an extra 75 thousand

cases amongst non-smokers against an extra 250 thousand cases amongst smokers. Note that this approach requires knowledge of the absolute risks, rather than just relative risks.

Another approach would be to use the *population* attributable fraction. The population attributable fraction (PAF) for exposure to a single factor is often interpreted as the fraction of disease in a population that might have been avoided if the exposure had not taken place (or, making some strong assumptions, if the exposure could be eliminated). For a single factor it can be calculated as

$$(\text{Risk in whole population} - \text{Risk in unexposed subgroup}) \div \text{Risk in whole population}$$

If smoking is the exposure of interest, the lung cancer risk in non-smokers is the average of that for non-smokers with both genotypes (since they have equal prevalence). So, using the information in Table 1, the population attributable fraction for smoking is  $(4.55 - 1.4) \div 4.55 = 0.692$  (69.2%). So 69.2% of the cases of lung cancer could have been avoided if cigarette smoking had not occurred.

The PAF is often used in a public health context to help decide which exposures to target. If it can be assumed that a number of different exposures all cause lung cancer, and there is an intervention to prevent their effect (e.g. eliminating an exposure or prophylactic treatment) then efforts may be directed towards whichever of exposures have the largest PAF. However, this approach is less useful if the intervention is not fully effective.

The PAF can be extended when there is more than one exposure category. This could be ordered categories of the same exposure factor (e.g. None, Low, Medium & High) or, as in our example, a combination of two factors (e.g. Unexposed & Low-risk-genotype, Exposed only, High-Risk-Genotype only, and High-Risk-Genotype & Exposed). The aim is to measure the effect on the population if the exposure-genotype combination in one subgroup had not occurred, using the doubly unexposed group (i.e. unexposed and low-risk genotype) as a reference group.

$$\text{Population attributable fraction (PAF)} = P_i \times (RR_i - 1) \div (1 + \sum P_i \times (RR_i - 1)) \quad ^2$$

- the subscript i refers to each of the three exposure combinations (++, +- & -+)

Using the information in Tables 1&2, for ++ subgroup

$$\begin{aligned} \text{PAF} &= 0.125 \times (13.33 - 1) \div (1 + 0.125 \times (13.33 - 1) + 0.125 \times (10 - 1) + 0.375 \times (1.33 - 1)) \\ &= 0.407 \text{ (40.7\%)} \end{aligned}$$

Values of PAF for the other subgroups are given in Table 2. This shows that the largest population impact comes from the Exposed & High-Risk-genotype subgroup. So, if intervention were feasible, there would be more benefit to the population as a whole in targeting the smokers & GSTM1-null subgroup.

Rather than look at the effect on the whole population, another approach is to use the *exposed* attributable fraction (EAF). The exposed attributable fraction for a single factor is the fraction of disease amongst the exposed subgroup that might have been avoided if that exposure had not occurred (or its effects could be eliminated). PAF is

used more widely, but the EAF is included here to show the distinction between them. For a single factor it can be calculated as

$$(\text{Risk in exposed subgroup} - \text{Risk in unexposed subgroup}) \div \text{Risk in exposed subgroup}$$

So, using the example in Table 1, the exposed attributable fraction for the single factor smoking (irrespective of genotype) is  $(14 - 1.4) \div 14 = 0.9$  (90%). This approach can be also be used when there is more than one exposure category: in our example the Non-smoker & GSTM1-wild subgroup is used as the 'unexposed' subgroup. The results are shown in Table 2. The subgroup with the largest population impact with this measure is still the Smoker & GSTM1-null genotype, but using this measure, the impact is only slightly greater than that of the Smoker & GSTM1-wild genotype subgroup.

All three measures of population impact indicate that the Unexposed & High-risk-genotype subgroup has the largest impact (to a greater or lesser extent) in the numerical example. However, there are very real practical difficulties in acting on this. Firstly, to identify this subgroup, the population members with the high-risk genotype would have to be identified by genetic screening and their environmental exposure determined. Secondly, having identified this subgroup, it may be difficult or impossible to reduce their cancer risks down to those experienced by the Unexposed & Low-risk genotype subgroup. It may not be possible to remove the environmental exposure from those so identified (e.g. smoking cessation programs are only partially successful, and then annual risks of lung cancer in ex-smokers take some years to be reduced to those of never-smokers). The example used also has the feature that the high-risk genotype increases lung cancer risk on its own. Even if the effects of the environmental exposure were eliminated, that would still only reduce the risk to that of the Unexposed & High-Risk-genotype subgroup: there may or may not be interventions that could reduce their risk further (e.g. prophylactic treatment). Unless interventions to reduce cancer risk in a genetic subgroup exist, there is no point in genetic screening.

If an effective intervention is possible, then a final, more recent, measure of population impact is the Number Needed to Screen<sup>3</sup>. This combines together the prevalence of the genotypes, the risks of cancer in each subgroup and the reduction in risks which could be achieved if screening identified individuals where intervention was needed. To provide comparability with the other measures discussed above, assume that lung cancer risks in smokers could be reduced to those of non-smokers by some treatment or intervention (e.g. via smoking cessation program and/or chemoprevention). If we consider smokers with the *high*-risk genotype, they have a lifetime lung cancer risk of 16% that could be reduced to 1.6% if they were identified. This gives a number needed to treat (NNT) of  $1/(0.16 - 0.016) = 6.9$ . However only 50% have the high-risk genotype, so we have to screen  $6.9/0.5 = 13.8$  to prevent one case – so the Number Needed to Screen (NNS) is 13.8. A similar argument applies to the smokers with *low*-risk genotype giving a NNS= 18.5. With these assumptions there is little benefit to screening, since there is little difference between NNS for the two genotypes. However, a more reasonable assumption might be that lifetime lung cancer risk in smokers could only be reduced to a fraction of current levels, rather than down to the level experienced by non-smokers. A recent paper<sup>3</sup> used the NNS approach when assuming it was possible to reduce lifetime lung cancer risks in

smokers by 50% (rather than to the level of non-smokers), and also concluded that there was little advantage to screening for GSTM1.

In the numerical example used here, there is limited benefit in screening for the high-risk genotype for any measure of population impact. However, since smoking is a lifestyle choice that considerably increases the risk of lung cancer (and many other diseases), it is extremely unlikely in practice that genetic screening would be considered: the obvious approach would be target all smokers to reduce their smoking. Screening is more likely to be considered if the exposure is involuntary (e.g. exposure to an industrial chemical or family history of cancer). Even then, there are ethical, legal and social issues to be considered.<sup>4</sup> It should be noted that all these calculations are sensitive to changes in any of the estimates of risks, prevalence and interaction. Given the sample size requirements in genotype-environment investigations, it is rare to have precise estimates of all of these.

None of these measures of interaction or population impact can be interpreted in isolation. A recent paper has suggested a tabular layout that includes many of the measures discussed previously<sup>5</sup>. It would be helpful if there were consistent reporting of all the measures needed to interpret genotype environment interaction in future studies: it remains to be seen if this will happen.

## References

- 1) Ottman R (1996) Gene-environment interaction: definitions and study designs. *Preventive Medicine* v25 p764-770
- 2) Rockhill B et al (1998) Use and misuse of population attributable fractions. *American Journal of Public Health* v88 p15-19
- 3) Vineis P et al (2001) Misconceptions about the use of genetic tests in populations. *Lancet* v357 p709-711
- 4) Coughlin, SS et al (1999) BRAC1 and BRAC2 gene mutations and risk of breast cancer : public health perspectives. *American Journal of Preventive Medicine* v16 p91-98
- 5) Botto LD & Khoury MJ (2001) Facing the challenge of gene-environment interaction : the two-by-four table and beyond. *American Journal of Epidemiology* v153 p1016-1020