

**COMMITTEE ON CARCINOGENICITY OF CHEMICALS IN FOOD,  
CONSUMER PRODUCTS AND THE ENVIRONMENT.****CRITERIA FOR THE DESIGN OF GENE-ENVIRONMENT  
EPIDEMIOLOGY STUDIES****1. Introduction**

1. As discussed in CC/2000/07 and Annex 1 therein, there is increasing evidence that many cancers are a consequence of gene-environment interactions<sup>1</sup> (GEIs). Whilst environmental agents are responsible for initiation and subsequent progression of the oncogenic process, genetic differences amongst individuals in one or more of the effectors (e.g. enzymes, receptors) involved serve as predisposing or susceptibility factors. Considerable effort has been devoted to identifying specific genetic factors that affect individual susceptibility to cancer, particularly of an environmental origin. The completion of the Human Genome Project, and the rapid pace of the various SNP (single nucleotide polymorphism) mapping programmes are providing considerable additional impetus to further studies of this nature.

2. Genetic susceptibility (or genetic predisposition) can be defined as an increased likelihood of developing cancer as a consequence of a genetic mutation, which may or may not result in actual development of cancer in a given individual. This definition is important in that it specifies a mechanistic link between the mutation and cancer development. Implicit in the definition is the concept of penetrance, in that not all subjects harbouring the mutation will necessarily develop cancer. Whilst this may be because of differences in exposure, with respect to penetrance it is because of the involvement of additional factors (environmental and or constitutional/genetic).

3. Penetrance can be defined as the ratio of individuals who carry an allele predisposing to cancer and who develop cancer versus those who carry the same allele but who do not develop cancer, i.e. the likelihood that a person carrying a specific mutation will develop cancer, following exposure to an agent of concern, after adjustment for age if necessary. Cancer can be associated with high penetrance or low penetrance alleles. Penetrance can also be described as complete or incomplete.

4. In general, genes of high penetrance are affected only rarely within the population, they usually impact on the health or lifespan of homozygous affected individuals and, when rare, are described as inborn errors of metabolism or rare genetic traits. Cancers caused by high penetrance alleles appear most often in related individuals and exhibit an hereditary pattern. Hence, such alleles are easy to identify

---

<sup>1</sup> A gene-environment interaction can be defined as the occurrence of a combination of a genetic factor and an environmental exposure such that the relative risk of the combination is greater than the risk of the individual factors alone.

by conventional familial studies. For the above reasons, cancer due to gene-environment interactions rarely if ever involves high penetrance alleles.

5. Genes of low penetrance are often affected relatively frequently within the population and do not normally affect the lifespan or health of the individual. They will cause an increased predisposition towards cancer but are usually not sufficient to cause cancer. This is because the influence of other factors is necessary which, in combination, leads to the onset of tumour development. When the frequency of the less common allele of a gene is  $> 1\%$  (1,2) it is described as genetically polymorphic. Hence, a formal definition of a genetic polymorphism is: “a monogenic trait that is caused by the presence in the population of more than one allele at the same locus, resulting in more than one phenotype within the population, the less common allele occurring in  $>1\%$  of individuals”.

6. One of the main objectives of gene-environment studies of cancer is to identify subpopulations who are potentially at increased risk of developing cancer following exposure to a particular environmental agent as a consequence a genetic polymorphism, relative to the other genotype(s). Once identified, there is currently no consensus on how to include this information in risk assessment, or in risk management. GEI studies may also provide mechanistic information, helping implicate a specific environmental factor in causing a particular type of cancer in exposed populations.

7. A number of published GEI studies suffer from flaws in design and/or interpretation, reducing their potential value in cancer risk assessment. Limitations of published studies include lack of objectivity, absence of clear *a priori* hypotheses, weaknesses in delineation of case and control groups, methodological problems in determining genotype or phenotype, failure to take account of major potential confounders, poor quantitative summary of results, *post hoc* sub-group analysis, inadequate power, particularly when stratifying the population, and unexplained inconsistency,

8. Hence, there is a need for a set of criteria by which to assess the quality and significance of GEI studies of cancer. This and the accompanying two papers (CC/01/4 and CC/01/5) address common issues on the design, interpretation and regulatory implications of such studies. Whilst the immediate objective is to assist in the analysis of published papers, a longer term goal is to contribute to the debate on the correct design of such studies in the future.

9. Assessment of epidemiological studies for possible cause and effect inferences still relies on the criteria proposed by Bradford-Hill in 1965 (strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment, analogy). Whilst ideally, all of these criteria would be met, it has been argued that if the more critical of these criteria are met, i.e. strength of association, temporal relationship, coherence, consistency and, if possible, biological plausibility this would be sufficient to establish a strong case for cause and effect. The application of

Bradford-Hill criteria in the interpretation of gene-environment interactions is discussed in detail in the third paper in this series (CC/01/5).

10. Whilst the Bradford-Hill criteria are often applied retrospectively in epidemiological studies, clearly it would be possible to incorporate the testability of at least some of the criteria into study design. Some would almost be included as a matter of course, e.g. exposure assessment (biological gradient) and strength of association but others may not (e.g. temporality of effect relative to exposure and consistency through comparability of protocol, etc).

## 2. Types of study design

### 2.1. Gene characterisation studies

11. The recognition that many cancer susceptibility genes are likely to be low penetrance has led to the evolution of two major study designs to the assessment of gene-environment interactions. These are 1) epidemiological studies of candidate susceptibility genes (gene characterisation studies) and 2) genetic association studies (gene discovery studies). In the first, the influence of known polymorphisms 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 polymorphic loci. Most gene-environment studies 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. On the other hand, gene discovery studies are the most common means of identifying high penetrance cancer susceptibility genes.

12. Gene characterisation studies involve the *a priori* selection of candidate genes to be included in the protocol. Gene selection is based either on previous evidence of association with cancer or by virtue of knowledge (or likelihood) of some function that its causally related to cancer, e.g. carcinogen activation (see CC/01/4). It has been suggested that the most appropriate endpoint for such studies is censored age-at-onset tumour data (i.e. age-related cancer incidence on a specified date (censoring date), to allow for effects on time to cancer appearance). In designing a suitable study, consideration needs to be given to the mode of inheritance, the number of alleles that can determine phenotype (see para 44), the frequency of the alleles within the population, the strength of effect of the gene, the modifying effects of other gene or environmental influences (which may confound the study) and the nature of the genetic hypothesis to be tested – is a single locus to be tested or multiple loci. Many of these considerations will impact on the power of the study.

13. Most study designs for gene-environment interactions in cancer causation are variants of the common epidemiological cohort and case-control designs for assessing cancer risk in relation to measurable risk factors, here exposure and genotype. Both main designs have their strengths and weaknesses. Obviously, as cancer is usually a late-onset disease, cohorts normally require prolonged periods of follow-up and may

also require large groups. Nevertheless this study design is being used, for example in the EPIC (European Prospective Investigation into Cancer and Nutrition) study (3). A potential problem with case-control studies is that for low frequency polymorphisms, the genotype of interest may occur in only a very few subjects, even if it is a risk factor. These limitations may require the use of modified strategies, such as some form of multistage sampling (see para 21).

14. In **case-control studies** a group of case patients with the cancer of concern is compared with a group of control subjects, who are cancer-free at the time of study, in terms of potentially causal genetic and exposure factors. The groups are matched as far as possible for confounding factors, such as age, sex, race, which are not the primary focus of the study. In the most common GEI studies, **unrelated subjects** are used as controls. They may be population-based (often identified by GP number in the UK), hospital-based or spousal (accepting that sex will not be matched, but often allowing good matching for the exposure of concern). Unexposed subjects without the susceptibility genotype<sup>2</sup> can be used as the referent group, enabling comparison of all other groups under a variety of models (additive, multiplicative). The power of studies with this design is influenced by the frequency of the exposure of concern and of the susceptible genotype. It has been suggested that both need to be relatively common (i.e. >25%) for case-control studies to detect GEIs with any confidence (see 4).

15. In a variant of the most commonly used case-control design, unaffected relatives of cases are used as the control subjects (**case-related control design**). To date, this design has not been used to any extent in GEI studies of cancer. The potential advantage is that it provides unbiased, consistent estimates, even when risk factors correlate between relatives. However, bias can occur when there is a correlation between measured and unmeasured risk factors in matched case-control pairs. Variations of this design include the nature of the relatives used. Often, randomly selected, unaffected siblings or cousins are used, as these provide the best match for age. Alternatively, the parents can be selected. Here, one can compare the genotype of cases with that of a notional control, comprising the non-transmitted alleles of each parent. Whilst such designs can improve efficiency<sup>3</sup>, by reducing genetic variation between cases and controls, and possibly also oversampling for a rare genotype, there may be a loss in efficiency if exposure is overmatched.

16. In a **cohort study** a group of cancer-free individuals is identified, perhaps on the basis of exposure to a risk factor of concern, and then followed up over time to determine eventual cancer incidence in exposed and unexposed sub-groups of the

---

<sup>2</sup> In this and the following papers (CC/01/4 and CC/01/5) the term “susceptibility genotype” is used to indicate the genotype(s) actually or potentially at greater risk following exposure to an environmental agent. Note that for convenience it will include heterozygotes, where appropriate (i.e. when they are also at increased risk relative to unaffected (“wild type”) homozygous subjects).

<sup>3</sup> Efficiency can be loosely defined as the power to detect a change of a given magnitude. More precisely, it is the precision of the parameter estimates for a particular study design using a particular analysis. It can be viewed as the ratio of sample sizes required to achieve the same degree of statistical precision, and is closely related to the power of the analysis to reject the null hypothesis. Often efficiency is expressed in terms of the asymptotic relative efficiency

population. In a gene-environment interaction study, all of the subjects could be exposed to a particular environmental risk factor, and comparison would be of eventual cancer incidence between different genotypes of interest. One of the advantages of a cohort design is that, unlike in case-control studies, subjects can be studied for all possible effects (here cancer types) of the exposure of interest (see Lanholz *et al* (5) for a list of ongoing cohort cancer studies).

17. The need for the long follow-up necessary in such a design can be overcome by using a **retrospective cohort**. However, for gene-environment interaction studies, this would mean that suitable biological samples would have to have been archived such that genotyping of all members of the cohort could be performed. Whilst this has been increasingly included as part of the protocol of such studies over the last few years, a number of potentially useful retrospective cohorts cannot be studied due to the lack of such samples. In addition, there may be ethical issues in the use of archived samples from living subjects for purposes for which they were not originally obtained.

18. One of the most efficient designs is a **nested case-control study from an established cohort**. In this design, once sufficient cases have accrued within the cohort, appropriate controls are selected from the remainder of the cohort. It would be necessary to genotype only those cases and controls selected for the nested case-control component of the investigation. Once a cohort is established, a study of genetic factors can be achieved much more quickly than would otherwise be the case. Exposure assessment could be performed at various times after establishing the cohort, and hence would not be subject to recall bias (see para 47). This design is also well suited to estimation of penetrance of the gene of interest, as the whole study population will have been enumerated.

19. Another study design that could be used is the **case-only design**. Here, only cases with the cancer of interest are identified. Cases are assessed for an exposure of concern and those without the putative susceptibility genotype serve as the “control” group. Non-exposed cases serve as the referent group. To obtain valid estimates of the gene-environment interaction, exposure and the genetic factor must occur independently and the type of cancer should have a low occurrence in the general population (to avoid a high chance occurrence). Case-only studies offer greater precision than case-control studies, and provide comparable power for detecting gene-environment interactions to that of detecting the main effect in a case-control study. However, such studies do not permit estimation of the main effects of the genetic and environmental factors, and these must occur independently. Further, this design cannot detect gene-environment interaction models that depart from additivity.

20. A modification of this design allows estimation of the main effects. The so-called **incomplete-data-case-control design**, involves collection of information on both genotype and environmental exposure in the cases but on only one of these in the controls. Main effects can be estimated if genotype and environmental exposure are independent, and if the cancer of concern is rare. To ensure independence of genotype and environment exposure, it may be necessary to collect such information

on a random subset of controls, which may make the gain in efficiency of this design questionable compared to a full case-control study.

21. Various **multi-stage designs** are available to overcome the limitations of conventional case-control studies when either the exposure or the susceptibility genotype is rare. In these designs, the numbers of cases and/or controls with the rare factor of interest are increased in some way. **Countermatching** can be used as a method of sampling controls from a cohort for nested case-control studies. One way of achieving this would be to use a surrogate of the susceptibility genotype itself, such as family history. This would be appropriate when genotyping of the entire cohort was not deemed feasible. Obviously, the gain in efficiency will be determined by how predictive family history is of the genotype of interest. One could also countermatch for a rare exposure of interest. It has been estimated that this can increase efficiency of main effect estimation by approx. 25% (see 4).

22. In a **balanced design** study, rather than select cases and controls at random, subjects are selected on the basis of susceptibility genotype or exposure of interest. Hence, stage II of a **two-stage design**, would be to select all case subjects with a rare factor of interest (here genotype or environmental exposure – “exposed”), and then to identify the same number of subjects in the other three categories (“non-exposed” cases, “exposed” and “non-exposed” controls). The oversampling is taken into account in the analysis of the study. Balanced designs are more efficient than conventional case-control designs. In a two-stage design, it is assumed that the rare “exposure” is known for the entire population in stage I. This might require considerable effort, for example to identify a rare genotype in a large population of cases and controls. To avoid this, a **three-stage design** could be adopted. Here, stage I would comprise selecting all cases on the basis of a simply assessed surrogate for the factor of concern, and sampling the same number of “unexposed” cases and “exposed” and “non-exposed” controls. In stage II, specific factor assessment would then be performed on this subset. Stage III would involve selecting all cases “exposed” to the rare factor and sampling equal numbers in the other three groups. One of the main difficulties is identifying adequate surrogates for stage I sampling. Often the choice of such a design will depend on the feasibility of assessing the entire population for exposure and genotype of interest.

23. Characteristics of some of these study designs are summarised in Table 1.

24. The choice of study design will require consideration of a number of issues, including the nature of the outcome to be assessed (usually censored age-at-onset data), the mode of inheritance of the genes of interest; the number of mutant alleles at each gene locus (e.g. single mutant allele, highly polymorphic); frequency of the mutant alleles within the population; if known, whether the effect of the mutation is strong or weak; whether there are other genetic or environmental factors with strong effects (e.g. in breast cancer family history and estrogenic status such as age at menarche are known to be important risk factors); how many genes might be involved in determining susceptibility (is only a single polymorphism or are multiple polymorphisms to be studied). Study design will also depend upon statistical

considerations (see para 53 *et seq*), feasibility (e.g. can sufficient suitable controls be identified), likelihood of cooperation of subjects, cost-efficiency and resource requirements and the potential for bias through selective survival in cases and controls. Unfortunately, there has been little systematic comparison of the full range of study designs across the spectrum of objectives and possible scenarios that exist. The availability of techniques such as clinical trial simulation (6) might help in some of the choices involved.

## 2.2. Gene discovery studies

25. Until recently, gene discovery designs have not been used widely in GEI studies of cancer. This is because of the impracticalities involved in screening the very large numbers of subjects that would be necessary to detect genes 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 a study.

26. As in other gene discovery initiatives, there is considerable interest in the use of intermediate (surrogate) endpoints in such studies rather than cancer. There are several reasons for this, perhaps the most important of which is that such an endpoint should be closer to the responsible gene. This is particularly relevant in studies of cancer, as malignancy is a multi-stage, multi-factorial process, and given the low penetrance of most susceptibility genes, is likely to be influenced by several genes, environmental and other factors. For example, if a carcinogen is activated by a polymorphic enzyme of xenobiotic metabolism, there is a greater chance of observing a relationship between the respective genotype and mutation at a target site (e.g. susceptible guanine residues in a gene known to be mutable by the agent from studies *in vitro*) than with frank cancer burden. The use of a surrogate also reduces the follow-up necessary in cohort studies. The major caveat, of course, is that any intermediate endpoint used in this way should be proximal to and causal in tumour development.

27. Conventional linkage analysis has insufficient power to detect genes with a penetrance as low as that often encountered in GEI studies (see 7 for discussion of study designs). Risch and Merikangas (8) have estimated that even under ideal conditions, it would require 2500 families to detect a genotype relative risk of 2 by conventional linkage analysis. The possibility of performing genome wide, or other large scale genomic scans using high throughput genotyping approaches for 10's or even 100's of thousands of SNPs has been heralded as the solution to identifying the basis of genetic susceptibility to environmental causes of cancer. There are many assumptions in this that need serious consideration. These include issues of statistical power, analytical methodology, public health consequences (especially, as is likely, many genotypes, at a number of different loci, are each shown to confer a small increase in risk) and the resources necessary. However, these issues will not be addressed further in this paper. Much of the remainder of this paper relates to gene characterisation studies, though some comments will also apply to gene discovery studies.

### 3. Selection of genetic factors

28. In gene characterisation studies of gene-environment interactions, the choice of genetic factor(s) to study is crucial. The nature of possible candidate genes is discussed in paper CC/01/4 and will not be considered in detail here. Ideally, the gene of interest should have some known or plausible functional relationship to the exposure factor at issue. For example, an enzyme of xenobiotic metabolism should be expressed in a tissue that could have relevance to the site of occurrence of the tumour of concern. In part, this is related to the hypothesis to be tested, and discussed further below (see para 33). An illustrative example is CYP2D6 and lung cancer from cigarette smoking. The causal relationship between this enzyme and the etiologic agent was not clearly defined, and most hypotheses (primarily that it was involved in the metabolism of a carcinogenic agent) have been shown to be incorrect. However, perhaps the greater question is why were these studies performed in the first place, in the absence of a clearly delineated biological hypothesis.

29. Obviously, any candidate gene to be studied for interaction with environmental exposure as a risk factor in cancer should exhibit variation within the population. As indicated above, preferably the gene should be involved in a process known to be related to carcinogenesis. However, the list of possible processes is increasing rapidly (see CC/01/4) and includes carcinogen disposition, DNA repair, chromosomal stability, oncogene and tumour suppressor gene pathways, apoptosis, immune function and surveillance, cell cycle control, signal transduction, hormonal metabolism and effects, vitamin metabolism and effects, telomerases, obesity, nutrition, behaviour (e.g. substance-seeking) and neurotransmission. Also, a number of less well-studied genetic mechanisms may be involved, such as imprinting, non-chromosomal inheritance, epigenetic mechanisms (e.g. methylation) and transgenerational effects. Genetic traits should exhibit Mendelian inheritance.

30. Whilst SNPs are invaluable in the genomic localisation of genes of concern, increasingly SNPs in coding regions (cSNPs, estimated at  $\sim 5-10 \times 10^4$ ) or in potential regulatory regions (intronic, upstream and downstream non-coding regions – perigenic or pSNPs, estimated at  $\sim 2-5 \times 10^5$ ) and even in intervening stretches of DNA between genes (often called “junk” DNA – intergenic or iSNPs  $\sim 2 \times 10^6$ ) are being regarded as potential candidate polymorphisms in susceptibility determination in cancer. In principle, there is no reason why such a polymorphism is not a determinant of susceptibility. The problem lies in the way in which such polymorphisms are identified. Until recently, polymorphisms were identified on the basis of observing a phenotype and then establishing the molecular basis for this. Then, there was a clear relationship between genotype and phenotype (accepting that on occasion the molecular basis for a polymorphism is not initially correctly assigned, e.g. CYP2A6 (9)). In contrast, SNPs are identified simply as a consequence of resequencing genomic DNA from several individuals, on the basis of primary sequence comparison. Hence, here the phenotype (functional consequence), if any, of such a polymorphism will be completely unknown. In the case of a cSNP, the predicted consequence of the mutation on the amino acid sequence of the gene product may provide some indication of functional significance. However, this would need to be tested, for example by expression of the recombinant protein. In the case of pSNPs (and perhaps iSNPs), there could be an effect on transcriptional regulation

(and hence protein expression), but at present this could be determined only by experimentation. As an example of the difficulties involved, several studies have been performed on the association of CYP2E1 and CYP1A1 alleles characterised by SNPs, not always in the respective coding region, and tobacco-related lung cancer. However, at least some of the alleles in question result in no clear phenotype (10) (see also CC/01/5).

31. Candidate gene selection plays an important part in the application of the Bradford-Hill criteria (see CC/01/5) in establishing biological plausibility. Hence, from the foregoing it is apparent that with some GEI study designs, in which the biological consequences of a polymorphism are not known, and hence should be regarded as more association studies than characterisation studies, *a priori* it will not be possible to establish biological plausibility from the study itself. Whilst such a study design might provide some indication of the role of a particular genetic polymorphism in determining susceptibility, this should perhaps be regarded as unconfirmed, until either further information on the biological consequences of the polymorphism is obtained, or a repeat study confirms the observation.

32. The effects of polymorphism of some genes is likely to be more specific than of others. For example, polymorphism of an enzyme of xenobiotic metabolism should affect the risk only of those compounds metabolised by that enzyme whereas polymorphism of a DNA repair enzyme should impact on the risk of all carcinogens of which adducts are repaired by that enzyme. Hence, some consideration needs to be given to likely biological specificity of the candidate polymorphism to be studied, both with respect to the range of exposures that might be affected and the nature of the cancer that might occur.

## 4. Aspects of study design

### 4.1. Study objectives and hypothesis

33. It is well established in epidemiology that sound study design includes a clear statement *a priori* of the hypothesis to be tested. In the past, this has often been implicit, particularly when only a single combination of genetic and environmental exposure factors are being studied. However, as the number of polymorphisms included in a study increases the need for an explicit statement of hypothesis is much more important. This is compounded by the fact that many of the genes studied are of low, if any, penetrance, so that it is difficult to demonstrate a significant increase in risk based on a simple comparison between cases and controls. As a result, the study group is often sub-stratified, perhaps on the basis of exposure level, but often on the basis of other factors for which there is no *a priori* basis. The net effect is to weaken dramatically the power of the study. Indeed, it could be argued that *a posteriori* hypothesis testing can almost never be used as a basis for establishing a causal GEI in cancer. The investment in time and resources on a large scale GEI study is perhaps amongst the greatest in biomedical research. Hence, there is a quite understandable desire to obtain as much information as possible from the study. When this is at the expense of sound study design and valid conclusions, however, it becomes counter-productive. It is not difficult to construct a plausible basis for an *a posteriori*

hypothesis, and present it as an *a priori* one. However, the power of the study will be fatally compromised. Perhaps one solution is to request that the study hypothesis and objectives should be submitted to an independent third part such as one of the Royal Colleges or relevant learned society, prior to data analysis (a similar suggestion has been made by Cuzick (11)). For example, the Molecular Epidemiology Group (MEG), which is affiliated to the UK Environmental Mutagenicity Society (UKEMS) and which has a particular interest in promoting best practice in the field of gene-environment interaction studies, might be persuaded to use their good offices to act as “honest brokers” in this way.

#### **4.2. Selection and characterisation of cases and control**

34. It is essential that the study population is relevant to the risk assessment being undertaken and it should be representative of a well-defined underlying cohort or population at risk, to permit effective implementation of the findings. Cases need to be clearly defined, with inclusion and exclusion criteria agreed in advance and specified in the study protocol. Definition of cases requires diagnostic criteria. In the case of specific cancers this usually necessitates confirmation by histopathological examination of tumour tissue. Issues such as whether to include subjects with premalignant changes need to be resolved. Will age criteria be applied and if so what are the age limits for inclusion? Other criteria that need to be considered are likely exposure to the agent of concern (see para 37 *et seq*), completeness of data file on a subject (at what point will a subject be excluded on the basis of incomplete information, e.g. obviously in a GEI study, lack of information on genotype would be a problem). In case-control and related designs, cancer prevalence (i.e. the proportion of a population that are cases at a given point in time) is used. This requires that a specified date for case diagnosis be agreed, after which no new diagnoses will be accepted. In cohort studies, cancer incidence (i.e. the rate at which new cases occur in the population over a specified period) is more commonly used. Again, this requires agreement of a specified cut-off date (censoring date) for case accrument.

35. Cancer registries can be invaluable for obtaining accurate information on background rates of specific types of cancer for the total population or for sub-groups, identified on the basis of variables such as age, sex, region or ethnicity. A cancer registry collects comprehensive information on all new cases of cancer occurring in the defined population and also often collects information on cancer deaths in the same population. Information collected usually includes details of cancer site and morphology, patient treatment, definition of the population being monitored, the period covered and an accurate estimate of the size of the population being monitored. This last information is essential for effective use of the incidence rates obtained.

36. Controls must be appropriate to study design and the cases recruited (see para 34). Wherever possible, controls must be matched for potential confounding (see para 48). In addition, there should be a similar range of exposure to the environmental agent of concern as in the cases. The design will require controls of known relatedness to the cases, either independent (no relation) or related in a specified way (e.g. sibling). Consideration needs to be given as to whether and, if necessary, how such relatedness will be confirmed. Where controls are specified as a random

independent sample from the underlying population, the method or randomisation needs to be established, and adequate recruitment needs to be achieved to avoid possible bias (see para 46). The power of a study to detect a statistically significant result can be increased by selecting more controls than cases, usually two or three. There is little to gain in having more than 4 controls per case. Background demographic information on cases and controls will need to be collected, using a validated technique, such as structured interview or questionnaire. Both hospitalised and healthy controls have been used in GEI studies. Hospital controls allow greater control of the subjects and also there is generally a better response rate. The disadvantage is that the disease itself (e.g. patients with non-malignant lung disease may be used as controls for cases with lung cancer) may affect measurements of either exposure or genetic status (see para 37 *et seq*). There could also be a differential effect on survival.

### 4.3. Characterisation of exposure

37. In any GEI study, the assessment of exposure, in both cases and controls, is essential to the successful interpretation of the data. Ideally one would obtain accurate information on internal exposure to the agent of concern (e.g. by use of an appropriate biomarker – see para 40). However, much more likely, exposure will be assessed from information relating to environmental levels of the agent. Hence, data on the level, duration, route and frequency of exposure will be required. It may be possible to use this to estimate body burden, for example by simple kinetic modelling (such as by applying a correction value for dermal versus oral exposure). In reality, exposure is often estimated quite crudely, for example for occupational category, or from some other poorly-correlated surrogate (e.g. smoking status of partner (yes/no) as surrogate for ETS exposure). Some consideration needs to be given to how sensitive the study conclusions will be to uncertainties in exposure estimates.

38. Occupational exposure characterisation, based on environmental or personal monitoring can provide some of the most useful data for estimating individual exposure, particularly when occupation is the major source of such exposure. Indeed, such information can be amongst the most quantitative available. Comparison of the levels of exposure attained in such situations can be compared with general environmental exposure levels, and may provide evidence for an exposure gradient that could prove invaluable in the analysis of dose-response relationships (see CC/01/5).

39. Exposure should be estimated for the period most critical in cancer aetiology. For many carcinogens, this will be early in the natural history of a tumour, when initiation plays a key role. Hence, in a case-control study, it may be necessary for subjects to recall earlier exposure. This is, of course, subject to large and possibly differential, error (recall bias, see para 47). In a cohort study, exposure can be assessed at various times throughout the study and hence the relevant exposure period will be available at the time of analysis. This does rely on identification of the exposure of concern in advance, which is not always possible.

40. Exposure information can be obtained by questionnaire, in which case pertinent questions will need to be devised and validated for the exposure of concern, structured interview, where care must be taken to avoid leading questions, or environmental monitoring, particularly in cohort studies. It is also possible to use biomarkers of exposure, although only a few of these have been adequately validated for quantitative use. Nevertheless, they can be very effective in semi-quantitative confirmation of a more subjective evaluation. Examples of biomarkers that have been used in GEI studies include DNA adducts of genotoxic carcinogens with lymphocyte DNA, urinary cotinine and urinary aflatoxin B<sub>1</sub> adducts. Again, exposure should be assessed for a period relevant to the natural history of the cancer, and thus current biomarker levels in cases (and controls) may not be appropriate. In addition, care must be taken that the disease itself, or its treatment, does not affect assessment of exposure. This could well occur in the use of a biomarker to study prevalent cases, but even the use of a questionnaire or responses in an interview could be affected by the cancer or its treatment.

#### **4.4. Characterisation of genotype or phenotype**

41. Obviously, the value of a GEI study depends on the ability to obtain a reliable estimate of the genotype or phenotype of concern. Rapid advances in genomics and pharmacogenetics over the last few years have meant the efficiency of detecting known polymorphisms is improving rapidly. A range of techniques is available, choice depending on the nature of the mutation and the resources available. Reliable genotyping information can now be obtained from almost any biological sample, including frozen whole blood, dried blood, hair, saliva, ethanol- and formaldehyde-fixed pathology specimens. The sensitivity of methods has also improved considerably, so that only one to a few micrograms, or even less, of DNA are necessary.

42. Methods in current use for genotyping have been reviewed recently by Blömeke and Shields (12). They all involve polymerase chain amplification (PCR) to provide the necessary sensitivity. However, a variety of techniques have been used to detect and quantify the resultant fragments from mutant alleles. These include: restriction fragment length polymorphism (RFLP) analysis, allele-specific oligonucleotide hybridisation, denaturing gel gradient electrophoresis, heteroduplex analysis, single-strand conformational polymorphism (SSCP), allele-specific PCR, allele-specific oligonucleotide probes, oligomer hybridisation, oligonucleotide ligation assay and primer-guided incorporation methods (e.g. genetic bit analysis). The last two techniques do not require an electrophoretic step. Perhaps the most frequently used to date are RFLP analysis and allele-specific PCR. When using the former, it is important that the RFLP site corresponds to the mutation of interest, and is not simply in linkage disequilibrium. Oligonucleotide ligation and genetic bit analysis are well suited to automation and high throughput, and hence are particularly useful when analysing large numbers of samples. Alternatives include dot or slot blot techniques, which are particularly sensitive. Recently, it has been shown that mass spectrometric techniques can be used for rapid genotyping of samples (13).

43. The post-genomic era is resulting in renewed efforts to identify genetic factors in human disease. As indicated above, one of the main approaches that will be used will be association studies with dense SNP maps of the human genome, requiring genotyping of many thousands of alleles in large numbers of subjects. This will require novel approaches to high throughput genotyping, which will inevitably be applied to GEI studies. Even when studies focus on only a small number of polymorphisms, the improved sample throughput provided by these methods may prove advantageous. Techniques that are already in development include automated mass spectrometric analysis, often of tagged PCR products (e.g. see 14), and single base extension approaches such as SBE-TAGS (15).

44. As an alternative to genotyping, the phenotype of cases and controls can be determined. Classically, this method has been widely used, but is now being largely superseded by genotyping methods. Phenotyping requires the availability of a suitable marker of gene function that can be studied in living subjects, either by *in vivo* administration or on suitable tissue samples. A number of probe substrates have been identified for phenotyping enzymes of xenobiotic metabolism, particularly some of the P450 forms and N-acetyltransferase 2 (NAT2) (16,17). Such probes must be specific for the target enzyme and chemical analysis should be simple and robust. In free living controls or cancer cases there is a possibility of interference either with enzyme activity or with analyte determination through ingestion of dietary compounds, drugs and other substances. Whilst the study protocol can reduce this potential source of interference, it might not be possible to eliminate it completely. It is also possible that the cancer process itself will alter apparent phenotype (phenocopy). Another disadvantage of phenotyping is that it measures activity at the time of assessment, which may not be the same as at the time of critical exposure. However, at least phenotyping does provide a functional assessment of a polymorphic locus, unlike genotyping where functional inference must rely upon the results of studies designed for this purpose (see para 30).

#### **4.5. Duration of follow-up**

45. Whilst particularly relevant for cohort designs, consideration needs to be given to the natural history of the cancer of concern even in case-control and other cross-sectional designs. Hence, whilst the objective is to collect cases as rapidly as possible, sufficient time should have elapsed since the start of exposure for there to be a reasonable probability of cancer occurring. If too short, case accrual will be too low to provide adequate power. If too long, survival bias may result in selective depletion of the susceptible group. Hence, in a case-control study, duration of exposure should be an inclusion criterion. In a cohort study, there will be a minimum interval before a meaningful analysis can be undertaken.

#### **4.6. Consideration of bias and confounding factors**

46. Ideally, in a GEI study only exposure and genotype of interest would vary within the population. Obviously, this is almost never the case. Hence, it is important to identify potential sources of bias that might be inherent in the study design or data collection methods. Bias can be of various types: selection bias, information bias and

confounding. Selection bias occurs when the basis for inclusion of cases and controls differs such that the groups are not strictly comparable. It is much more of a problem, in case-control studies than in cohort studies, where all subjects are included. Selection bias in controls can be avoided by identifying a random sample and obtaining 100% response rate. In practice this is impractical for a number of reasons. Selection bias can be controlled for in the statistical analysis, in the same way as confounding. Sources of selection bias include lack of comparability between cases and controls in factors such as diet, lifestyle, socio-economic status, geographical location; differences in recall between cases and controls; differences in data collection such that there is unequal ascertainment of health effects; unequal follow-up of cases and controls. In a cancer study, by definition cases will not be healthy whereas controls often will be. This can be a major source of bias through indirect effects of the disease or its treatment on survival from non-cancer causes, participation rates, estimation of exposure (directly or indirectly) and assessment of phenotype. Only genotype assessment cannot be affected, but of course there could be selective depletion of one genotype due to differential survival. Where hospital controls are used, there can be selection bias through an effect of the polymorphism on the likelihood of developing the disease upon which inclusion in the control group is based.

47. Information bias has been described as bias amongst subjects who have been recruited into a study, whereas selection bias relates to subjects prior to inclusion in a study. Information bias includes issues of accuracy and validity of the methods used to assess exposure and genotype or phenotype (see elsewhere in this paper for details). For example, if exposure estimation relies upon information provided by the subject via questionnaire or interview, there is considerable potential for deliberate or unintentional inaccuracy of recollection of relevant details, so-called recall bias. The result of such information bias is misclassification. The frequency of the affected allele will affect the likely consequences of misclassification. When the allele is common, a small decrease in sensitivity (probability of correctly classifying subjects with the susceptibility genotype) may result in misclassification of a large number of subjects whilst a decrease in specificity (probability of correctly classifying subjects as without the susceptibility genotype) will have much less effect. When the allele is rare, the converse is true. Non-differential information bias (same likelihood of misclassification in cases and controls, or in exposed and non-exposed subjects) generally increases the risk of producing a false negative result, i.e. a relative risk estimate of 1.0. Differential information bias can increase or decrease the estimate of the relative risk from the unbiased estimate. The validity of a study can be severely compromised even by relatively small amounts of misclassification. The impact that this can have on power and study size is considered further below (see para 60).

48. A further possible source of bias is confounding. This refers to the non-comparability of sub-groups within the study population. A confounding variable is a risk factor, independent of those of concern (here exposure and genotype), that is unequally distributed amongst the study sub-populations such as the different genotypes. Examples would be cigarette smoking and lifestyle. In the absence of the counteractive effect of other biases, for confounding to occur a factor must influence cancer risk and it must be associated with either the genotype or environmental exposure of concern. Factors that are involved in the cancer process itself, between

initial exposure and final outcome, are not confounders, although this can be difficult to establish.

49. The nature of epidemiological studies (i.e. essentially observational) is such that it is often not possible to control in advance for such confounding variables, which may affect study outcome. Where possible, adjustment for potential confounding factors is achieved in the design of the study (e.g. by matching cases and controls for such factors). However, it is often necessary to adjust for confounding in the statistical analysis of the data.

50. Confounding is a particular problem in GEI studies because so many lifestyle factors tend to correlate with each other, for example smoking and alcohol ingestion. Hence, knowledge of which factors correlate with each other is required, together with their accurate assessment. Confounding can also occur at the genetic level, due to linkage disequilibrium. Here, alleles on one gene (associated with increased risk) are inherited with specific alleles on adjacent genes (unrelated to risk). Hence, analysis of the genotype of these genes will appear to be associated with increased risk. Thus, in the absence of knowledge of which alleles are co-inherited it is important to establish the mechanism of carcinogenesis of an environmental agent.

51. One way of controlling for confounding is to stratify the data according to the levels of the confounder(s) and then to calculate a summary effect estimate for the information in each sub-group. However, it is usually not possible to control for more than two or three confounders using this approach. Mathematical modelling may help, but here there is a potential problem of multicollinearity if variables are highly correlated. Confounding is best controlled by effective use of *a priori* knowledge, together with consideration of the extent to which the risk changes when the factor is controlled in the analysis. However in GEI studies, information on how genetic markers are associated with exposure is rarely available.

52. When it is not possible to control directly for a confounder, it might still be possible to obtain some indication of its strength of effect by, for example, the use of a surrogate measure. Alternatively, it may be possible to obtain information on the confounder in a sub-group of cases and controls, and assess the magnitude of the effect. In other approaches, such information is obtained on only a sub-group of controls or even on a group of comparable subjects from the study base.

#### **4.7. Power and sample size necessary to detect an effect**

53. Statistical power is a function of the prevalence of the at risk genotype and the magnitude of the risk. Hence, for polymorphisms with a large proportion of the susceptible genotype and a high relative risk, only small numbers of subjects are required, whereas large populations are required when the proportion of the susceptible genotype is small and the relative risk is low. In designing a GEI study, as in any epidemiological study, it is essential that power calculations are undertaken in advance to obtain a realistic estimate of the number of subjects necessary to ensure a statistically meaningful outcome. When multiple comparisons are to be undertaken,

for example by stratifying the subjects in various ways, or by analysing multiple polymorphisms, account should be taken of this by using the appropriate statistical approaches (see para 59).

54. Lubin and Gail (18) have developed a sophisticated mathematical approach to assess the power of a study or the number of subjects required in a GEI study. This is based on a standard multivariate logistic regression model of the form:

$$\text{Logit} [P(D = 1|E, G)] = \hat{\alpha}_0 + \hat{\alpha}_E E + \hat{\alpha}_G G + \hat{\alpha}_{EG} EG$$

where  $\beta_G = \ln(\text{OR}_{G=1|E=0})$ ,  $\beta_E = \ln(\text{OR}_{E=1|G=0}) = \ln(\text{OR}^{\text{tb}}_{E|G=0})$  and  $\beta_{EG} = \ln(\theta) = \ln(\theta^{\text{tb}})/(Q-1)$ .

The notation used is as follows: D, presence (1) or absence (0) of cancer; G, genotype (0,1 for non-susceptible and susceptible, respectively); E, exposure to environmental factor with Q levels of exposure with values of 0, 1, ... Q-1;  $\text{OR}_{G=1|E=0}$ , odds ratio for the gene effect in unexposed subjects (lowest exposure category);  $\beta_0$ , the baseline cancer rate;  $\text{OR}_{E=1|G=0}$ , odds ratio for environmental exposure of level 1 in genetically non-susceptible subjects; the odds ratio increases with increasing exposure as a power of  $\text{OR}_{E=1|G=0}$  (i.e. the log odds ratio of cancer is a linear function of exposure);  $(\text{OR}_{E=1|G=0})^{Q-1}$ , odds ratio for the highest to lowest (top-to-bottom) exposure levels in genetically non-susceptible subjects ( $\text{OR}^{\text{tb}}_{E|G=0}$ );  $\theta$ , effect of the gene-environment interaction for environmental exposure of level 1;  $(\theta)^{Q-1}$ , effect of the gene-environment interaction for the highest to lowest (top-to-bottom) exposure levels ( $\theta^{\text{tb}}$ ).

55. In a case control study of a GEI, the first stage in the statistical design is to specify an alternative hypothesis,  $H_A$ , which represents the “true state of nature”. This implies that all of the parameters in the model are specified, including the magnitude of the interaction to detect ( $\beta_{EG}$  or  $\theta^{\text{tb}}$ ), the odds ratio for the main effects ( $\text{OR}_{G=1|E=0}$  or G,  $\exp(\beta_E)$  and  $\text{OR}^{\text{tb}}_{E|G=0} = \text{OR}_{E=Q-1|G=0}$  or E,  $\exp(\beta_E)$ ).  $\beta_0$ , baseline cancer incidence (in genetically non-susceptible, non-exposed subjects) must also be specified. However, if the incidence of the cancer is rare, the magnitude of this term will have little effect on sample size.

56. The next step is to specify the null hypothesis  $H_0$ . To test for the absence of a multiplicative GEI (i.e. the OR for exposed susceptible subjects is simply the product of the individual factor-specific ORs), the null hypothesis is designated  $H_0$ :  $\beta_{EG} = 0$ , equivalent to  $H_0$ :  $\theta^{\text{tb}} = 1$ .  $H_A$  is defined by  $\beta_0$ ,  $\beta_G$ ,  $\beta_E$  and  $\beta_{EG}$  whilst  $H_0$  is defined by  $\beta_0$ ,  $\beta_G$  and  $\beta_E$  with  $\beta_{EG} = 0$ . The maximum likelihood estimates of  $\beta_G$  and  $\beta_E$  can be used when  $\beta_{EG} = 0$  and the alternative is true. To obtain the most accurate estimate of power and sample size when testing for  $\beta_{EG} = 0$ , it is necessary to use the most likely values of  $\beta_G$  and  $\beta_E$  for when  $H_A$  is true (see 19).

57. Table 2 illustrates the effects of penetrance (which here can be estimated as the odds ratio for the effect of genotype on exposed subjects,  $\theta^{\text{tb}}$  if genotype is not a risk factor in the absence of exposure) and frequency of the susceptibility genotype on

sample size to detect a simple two-factor gene-environment interaction. It is apparent that for ORs of the magnitude often encountered in GEI studies (approx. 2), even with relatively common polymorphisms (i.e. susceptible genotype ~ 50%, e.g. GSTM1), over 1000 cases would be required to detect a significant interaction. As the frequency of the susceptible genotype decreases, the number of subjects required increases considerably. Even with an OR for the interaction of 5, the number of cases required with a susceptibility genotype frequency of 5% is very large. Whilst modern genotyping methods will improve sample throughput, there are still other major resource implications, e.g. exposure assessment, in studying such numbers.

Table 2. Effect of penetrance and genotype frequency on sample size in GEI studies. The baseline cancer rate is 0.001 (for rare cancers, the actual rate has little impact on sample size). The OR for cancer from exposure alone (in non-susceptible subjects) is 1.5. 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 (19)).

Proportion of susceptible genotype in population	$\theta^{tb}$	Number of subjects (same number of cases and controls)
0.5 (50%)	2	2215
	5	485
0.2	2	3891
	5	1017
0.05	2	13902
	5	3949

58. In the above analysis, it is assumed that exposure and genotype are independent. If this is so, a case-only study would remove the control component of the variance and provide the same power as a study with a larger number of controls per case.

59. As indicated above, in large GEI studies, often more than one hypothesis is investigated, sometimes *a priori* but often *a posteriori*. Hypotheses may relate to risks in subjects stratified according to some potentially interacting variable such as ethnicity, sex, dietary habits, age, or to different genetic polymorphisms, alone or in combination. Traditionally, adjustment in such circumstances was by methods such as the Bonferroni correction. However, the current view is that this is too conservative, and that the resultant demands on statistical power were unrealistic. Whilst many studies of GEIs in cancer now include no correction for multiple comparisons, relying on current knowledge for interpretation, it has been argued that this also is not appropriate and that, at least in some circumstances, statistical allowance using Bayesian adjustments should be made. The main options are empirical Bayes (EB) and semi-Bayes adjustment.

60. Bayesian adjustments should be used when there are a large number of comparisons (particularly when there is an absence of prior knowledge), comparisons can be grouped, and within the groups the comparisons are similar, random error makes a large contribution, or some choice of follow-up investigations needs to be

made. Even in other circumstances, Bayesian adjustment may be useful. EB produces more accurate estimates of relative risk, and hence permits more effective public health measures, as resources are finite. The methodology involved is described in detail by Steenland *et al* (20). These authors suggest that it may be appropriate to present both unadjusted and Bayesian-adjusted findings, as an aid to decision-making regarding potential further work.

61. Whilst it is often assumed that cancer risk in the susceptible genotype relative to that in the non-susceptible genotype increases as a function of exposure, this may not always be the case (see CC/01/4). For example, it may be that the basis of genetic susceptibility is a reduced capacity to detoxify a carcinogen. At high doses, detoxication could become saturated in both susceptible and non-susceptible genotypes, so that there is now little difference in relative risk between the groups (low exposure gene (LEG) effect). Taioli *et al* (21) have devised a standard logistical regression approach to test for the nature of the interaction, particularly whether there is a high exposure gene (HEG) effect or an LEG effect. One issue yet to be resolved is the impact that such *post hoc* analysis has on statistical power. Clearly the power of the study is reduced, but it is difficult to establish by how much.

#### **4.8. Methodology for data collection and analysis**

62. Many of the relevant issues have been discussed above. There should be a clear study plan, with specific protocols for each phase. The information required from each aspect of the study (e.g. genotype, diagnosis, exposure assessment, demographic data) should be stated explicitly, and where possible should be recorded on pro-forma designed for the purpose. Exclusion and inclusion criteria need to be stated explicitly and known to those recruiting subjects. Response rates must be recorded, together with the procedures to be followed in the event of a negative response. The number and nature of the analyses to be performed should be clearly stated in advance. In studies including cases and controls, there should be independent coding to blind those involved in the analyses. Similarly, where possible, statistical analyses should be performed blind to the nature of the sub-groups (e.g. genotype identity, exposure classification). A strategy for the statistical analysis of the data should be devised, and criteria for deviating from this established. If any *a posteriori* hypothesis testing is to be undertaken the basis for this must be agreed in advance, at least in general, and a separation between such analyses and those testing *a priori* hypotheses maintained.

63. Procedures for quality control should be established for all major aspects of study design and implementation (e.g. identification and selection of subjects, data collection methods, exposure assessment, genotype determination, statistical analyses).

#### **4.9. Response rate and handling missing data**

64. As in other epidemiological studies, it is important that the response rate is recorded accurately, together with the reasons for lack of response, if possible. Low

response rates complicate subject recruitment to the target necessary to achieve the power required. Even more importantly, they introduce a potential source of bias which could, in some circumstances threaten the validity of the entire study. This would be particularly true if the response rate was selectively low in one sub-group, for example in those with the susceptibility genotype. There are many possible reasons for this, for example if using a phenotyping test, one genotype might be reluctant to participate due to prior experience with the probe substrate. Alternatively, the genotype might be associated with a disease other than cancer. In assessing the significance of response rate, it is important to consider how unrepresentative the non-responders are relative to the objectives of the study. Of course, this might be difficult to assess. To assess the likely bias, a small random sample of the non-responders could be identified and every effort made to encourage their participation. Information from this sub-group should help establish the extent of bias among other non-responders. An alternative strategy is to use what information is available for subjects in the entire study base, and to compare the profile of responders and non-responders for important characteristics, such as age, sex and socio-economic status. One way of evaluating the maximum possible impact of the bias introduced by the response rate, is to assume the worst, for example that all of the non-responders had the susceptible (or the non-susceptible) genotype and determining the effect that this has on the statistical outcome of the study. This is a form of sensitivity analysis.

#### **4.10. Reporting of data**

65. The study should be reported in adequate detail, providing a transparent description of the methods used, particularly for the statistical analyses. Full details of potential bias and confounding should be presented, together with information on how their possible impact was minimised. The cases and controls should be described, together with inclusion and exclusion criteria. The methods used for genotyping or phenotyping subjects should be described in sufficient detail for their reliability to be assessed. If sensitivity analysis<sup>4</sup> was undertaken, the results should be presented. The hypotheses tested should be reported, and whether these were *a priori* or *a posteriori*. Any correction for multiple comparisons should be discussed. The extent to which the conclusions fulfil the Bradford-Hill criteria should be discussed (see CC/01/5). The relative risks for each of the major effects, together with their confidence intervals, should be clearly stated. Where appropriate, the population attributable risk should be reported (see CC/01/5).

### **5. Conclusions and recommendations**

66. A range of study designs is available for performing gene-environment interaction studies. However, the most efficient designs have yet to be determined by systematic comparison of the available alternatives. This could be accomplished by simulation.

---

<sup>4</sup> Sensitivity analysis is an evaluation of the statistical outcome of the data analysis to the assumptions underlying the analysis, e.g. degree of confounding by a specific factor.

67. The hypotheses being tested in GEI studies are not always clearly stated. This is becoming increasingly important, as the number of comparisons undertaken in each study increases. Thought should be given to strengthening the robustness of the *a priori* hypotheses, perhaps by lodging them with a third party in advance of data analysis and at the very least stating them clearly in the study report.

68. The number of multiple comparisons undertaken in GEI studies is large and is increasing rapidly. Statistical analyses of the data often do not take this into account. Some work is necessary to establish which techniques are the most appropriate to analyse studies of this type.

69. Selection of candidate genes is becoming increasingly complex (see CC/01/4). Phenotyping has the advantage that it reflects gene function, whereas this is not necessarily the case for genotyping. However, phenotyping on the scale now necessary is impractical. Hence, the justification of gene selection should be based on biological plausibility, i.e. a demonstrable effect of a mutation on gene function.

## 6. References

1. Jackson PR, Boobis AR, Tucker GT (1991). Phenotype or genotype. *Br. J. Clin. Pharmacol.* **31**, 119-120.
2. Lackie JM, Dow JAT (1999). In: *The Dictionary of Cell & Molecular Biology* (Third edition), Academic Press, London.
3. Riboli E (1992). Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann. Oncol.* **3**, 783-791.
4. Goldstein AM, Andrieu N (1999). Detection of interactions involving genes: available study designs. *Monogr. Natl. Cancer Inst.* **26**, 49-54.
5. Lanholz B, Rothman N, Wacholder S, Thomas DC (1999). Cohort studies for characterizing measure genes. *Monogr. Natl. Cancer Inst.* **26**, 39-42.
6. Bonate PL (2000). Clinical trial simulation in drug development. *Pharm. Res.* **17**, 252-256.
7. Schaid DJ, Buetow K, Weeks DE, Wijsman E, Guo S-W, Ott J, Dahl C (1999). Discovery of cancer susceptibility genes: study designs, analytic approaches, and trends in technology. *Monogr. Natl. Cancer Inst.* **26**, 1-16.
8. Risch N, Merikangas K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.
9. Oscarson M, McLellan RA, Gullsten H, Agundez JA, Benitez J, Rautio A, Raunio H, Pelkonen O, Ingelman-Sundberg M (1999). Identification and characterisation of novel polymorphisms in the CYP2A locus: implications for nicotine metabolism. *FEBS Lett.* **460**, 321-327.
10. Ingelman-Sundberg M. (1997). The Gerhard Zbinden Memorial Lecture. Genetic polymorphism of drug metabolizing enzymes. Implications for toxicity of drugs and other xenobiotics. *Arch Toxicol Suppl.* **19**, 3-13.
11. Cuzick J (1999). Interaction, subgroup analysis and sample size. In: *Metabolic Polymorphisms and Susceptibility to Cancer* (P Vineis, N Malats, M Lang, A d'Errico, N Caparaso, J Cuzick, P Boffetta, eds), IARC Scientific Publications No. 148, IARC, Lyon. pp. 109-121.

12. Blömeke B, Shields PG (1999). Laboratory methods for the determination of genetic polymorphisms in humans. In: *Metabolic Polymorphisms and Susceptibility to Cancer* (P Vineis, N Malats, M Lang, A d'Errico, N Caparaso, J Cuzick, P Boffetta, eds), IARC Scientific Publications No. 148, IARC, Lyon. pp. 133-147.
13. Jackson PE, Scholl PF, Groopman JD (2000). Mass spectrometry for genotyping: an emerging tool for molecular medicine. *Mol. Med. Today* **6**, 271-276.
14. Kokoris M, Dix K, Moynihan K, Mathis J, Erwin B, Grass P, Hines B, Duesterhoeft A (2000). High-throughput SNP genotyping with the masscode system. *Mol. Diagn.* **5**, 329-340.
15. Hirschhorn JN, Sklar P, Lindblad-Toh K, Lim YM, Ruiz-Gutierrez M, Bolk S, Langhorst B, Schaffner S, Winchester E, Lander ES (2000). SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. *Proc. Natl. Acad. Sci. USA* **97**, 12164-12169.
16. Kivisto KT, Kroemer HK (1997). Use of probe drugs as predictors of drug metabolism in humans. *J. Clin. Pharmacol.* **37** (1 Suppl), 40S-48S.
17. Streetman DS, Bertino JS, Nafziger AN. (2000). Phenotyping of drug-metabolizing enzymes in adults: a review of in-vivo cytochrome P450 phenotyping probes. *Pharmacogenetics* **10**, 187-216.
18. Lubin JH, Gail MH (1990). On power and sample size for studying features of the relative odds of disease. *Am. J. Epidemiol.* **131**, 552-566.
19. García-Closas M, Lubin JH (1999). Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am. J. Epidemiol.* **149**, 689-692.
20. Steenland K, Bray I, Greenland S, Boffetta P (2000). Empirical Bayes adjustments for multiple results in hypothesis-generating or surveillance studies. *Cancer Epidemiol. Biomarkers Prev.* **9**, 895-903.
21. Taioli E, Zocchetti C, Garte S (1998). Models of interaction between metabolic genes and environmental exposure in cancer susceptibility. *Environ. Health Perspect.* **106**, 67-70.