Biomarkers in Cancer Chemoprevention
Miller, A.B., Bartsch, H., Boffetta, P., Dragsted, L. and Vainio, H. eds
IARC Scientific Publications No. 154
International Agency for Research on Cancer, Lyon, 2001

Selection and validation of biomarkers for chemoprevention: the contribution of epidemiology

P. Vineis and F. Veglia

This chapter considers the epidemiological contribution of DNA adducts as an example of markers for use in chemoprevention studies, and highlights the potential biases inherent in the conduct of epidemiological studies with molecular markers. Although adducts have been interpreted mainly as biomarkers of exposure, 'bulky' DNA adducts such as those measured by ^{32}P-postlabelling or ELISA in white blood cells are more correctly interpretable as markers of cumulative unrepaired DNA damage. The latter concept can prove useful in cancer epidemiology, since it is consistent with existing knowledge on the importance of duration of exposure in the etiology of chemically-induced cancers. Increasing evidence suggests that in addition to prolonged exposure to genotoxic chemicals, inter-individual variability in carcinogen metabolism and DNA repair is predictive of cancer risk. Also from this point of view, measurements of 'bulky' DNA adducts can be useful as biomarkers for studies in populations, since they express the amount of carcinogen linked to DNA after repair, taking into account individual repair capacity. Finally, we suggest a theory of causality based on the work of Wesley Salmon and the concept of 'propagating mark', which is particularly attractive for molecular epidemiologists.

Categories of markers and their complex interplay

The usual categorization of biomarkers refers to three categories of markers, of exposure (internal dose), of intermediate effect and of individual susceptibility (see Chapter 1). We will examine a few examples that suggest that the nature of biological measurements may be more complex than is implied by such categorization.

There is increasing evidence that some types of DNA adduct can be considered both as markers of internal dose and as markers of susceptibility, because (*a*) they predict the onset of disease independently of exposure levels (see Example 1 below), and (*b*) they represent an integrated marker of exposure and of the individual's ability to metabolize carcinogens and repair DNA damage. Classical markers of susceptibility, such as genetically-based metabolic polymorphisms, can modulate the effect of exposures, but their expression can also be influenced by external determinants such as dietary habits. In this complex context of mutual interplay, the level of exposure is also likely to be important, as Example 2 demonstrates.

Validation

Validation of biomarkers can be interpreted in at least three different ways. One is the usual validation of a marker from the laboratory point of view, including an assessment of measurement error (for example by repeat measurements in the sample, and by the use of positive control samples with known values of the marker level). Measurement error is expressed by, for example, the coefficient of variation, the ratio between the standard deviation and the mean. Still in this category of validation measures, we can include measures of marker reliability (reproducibility and repeatability). A second category is represented by attempts to estimate the validity of the marker when compared with a standard (internal validity: sensitivity and specificity) and the impact of its use in a population (predictive value, depending on internal validity and the prevalence of the condition that we aim to measure). These aspects of validation are extensively dealt with in Toniolo *et al.* (1997).

A third concept of validation refers to the relevance of the marker in the context of population

studies or intervention trials when it is used as a surrogate end-point biomarker. Even if a marker is reliable and valid, and its predictive value reasonably high, in the context of chemoprevention we need to be sure that the marker is a good surrogate end-point. This is the category of markers we mainly refer to below.

Examples

Example 1: DNA adducts as predictors of cancer risk and dietary effect modifiers

'Bulky' DNA adducts (adducts with large molecules, as opposed to small alkyl radicals such as methyl groups) can be considered both as markers of internal dose and as markers of susceptibility. Several studies have considered the association between cancer at various sites and the levels of bulky DNA adducts (Table 1). Most studies (Tang *et al.*, 1995; Li *et al.*, 1996; Peluso *et al.*, 2000; Vulimiri *et al.*, 2000) have found that cancer cases had higher levels of adducts than non-cancer controls, after adjustment for relevant exposures such as smoking. This suggests that we may consider bulky DNA adducts (as measured in white blood cells by the ^{32}P-postlabelling method) as markers of susceptibility, in addition to their being markers of exposure. Further evidence for such an interpretation comes from a recent case–control study nested in the Physicians Health Study cohort, which found DNA adducts to be predictive of the cancer outcome (Perera, 2000). In the latter study, adducts were measured in blood samples that were collected years before cancer onset, thus ruling out the possibility that the higher adduct levels were due to metabolic changes associated with an already existing cancer.

It appears that bulky DNA adducts express cumulative exposure to aromatic compounds after the action of metabolizing enzymes and before the intervention of DNA repair enzymes. They are, therefore, markers of cumulative unrepaired DNA damage. There is a large amount of additional evidence to support this interpretation, based on the observation that the lymphocytes of cancer patients (and of their healthy relatives) have higher levels of DNA adducts after treatment with a hydrophilic chemical, compared with those of noncancerous individuals (Berwick & Vineis, 2000).

Other types of adduct provide a better expression of external exposure than of cumulative unrepaired DNA damage. This is the case for protein adducts that are not repaired.

The level of DNA adducts, which is suggested to be predictive of cancer risk, can be modulated by personal habits such as intake of fruit and vegetables. In a study of healthy volunteers, conducted within the EPIC Italian cohort, an inverse association between consumption of several dietary items and adduct levels was found (Table 1) (Palli *et al.*, 2000). Out of about 120 food items that were investigated, only those shown in this table were associated in a statistically significant manner with the adduct levels. Some of these food items, namely leafy vegetables and fruits, were associated with an approximately 25–30% decrease in adduct levels, while meat consumption was associated with a slight and non-significant increase (Palli *et al.*, 2000). When nutrients were assessed by means of food–nutrient conversion tables, associations were found with monounsaturated fatty acids and β-carotene (Table 2). Similar relationships have been observed in the investigation on bladder cancer by Peluso *et al.* (2000).

Other types of change that are related to DNA damage can be modified by fruit or vegetable intake. For example, several functional tests have been developed to explore individual DNA repair capacity (mutagen sensitivity tests; see Berwick & Vineis, 2000). The repair capacity thus assessed seems to be modified by several exposures or personal habits. In cultured lymphocytes, antioxidants such as α-tocopherol exercised a dose-dependent protective effect in preventing bleomycin-induced chromosome damage (Trizna *et al.*, 1992, 1993). In another study, Kucuk *et al.* (1995) found strong inverse correlations between plasma nutrients and results from the mutagen sensitivity assay. Correlation coefficients were 0.76 with β-carotene and 0.72 with total carotenoids. However, in randomized double-blind trials, Hu *et al.* (1996) and Goodman *et al.* (1998) did not find any association between supplementation and DNA repair activity.

From a mechanistic point of view, the modulating effect of fruit or vegetables on adduct levels may be explained by induction of enzymes involved in carcinogen detoxification or by repression of enzymes involved in carcinogen activation. As far as 'mutagen sensitivity' is concerned, this is usually interpreted as an indirect expression of

Table 1. Adjusted means[a] of relative adduct labelling (DNA adducts per 10^9 normal nucleotides) according to tertiles of daily consumption of selected food groups or food items, as reported on a food frequency questionnaire (EPIC Italy, 1993–98)

	Tertile of consumption				*p*-value
	Adjusted mean ± SE			% change[b]	for trend*
	I	II	III		
Leafy vegetables (except cabbage)	9.07±1.22	8.60±1.13	6.34±1.21	–30.10	**0.02**
Fruiting vegetables	8.34±1.34	8.48±1.14	7.29±1.19	–12.59	0.06
Root vegetables	7.60±1.14	10.29±1.12	5.89±1.21	–22.50	0.07
Cruciferous	10.22±1.2	16.57±1.15	7.42±1.15	–27.40	0.2
Grain and pod vegetables	7.16±1.20	9.51±1.23	7.65±1.16	6.84	0.7
Stalk vegetables, sprouts	8.76±1.16	8.70±1.17	6.66±1.17	–23.97	0.2
Mixed salad, mixed vegetables	8.47±1.19	8.45±1.18	7.02±1.33*	–17.12	0.2
All vegetables	8.92±1.27	8.66±1.13	6.58±1.18	–26.23	**0.01**
Legumes	8.24±1.22	8.18±1.16	7.69±1.15	–6.67	**0.01**
Potatoes	6.98±1.19	7.93±1.19	9.19±1.20	31.66	0.7
Onion, garlic	7.89±1.23	8.03±1.23	8.11±1.13	2.79	0.8
Fresh fruit (all types)	8.56±1.23	6.88±1.20	6.47±1.24*	–24.4	**0.04**
Nuts and seeds	6.54±1.38	7.79±1.19*	9.34±1.15*	42.81	0.1
Fruit and vegetable juices	9.34±1.10	6.40±1.43	7.42±1.22	–20.56	0.5
Milk	8.81±1.19	8.27±1.15	6.98±1.20	–20.77	0.8
Yoghurt	7.54±1.00	9.23±1.35	7.83±1.28	3.85	0.3
Cheese (including fresh cheeses)	7.02±1.20	6.86±1.14	10.20±1.20	45.30	0.1
Pasta, other grain	6.82±1.26	7.37±1.13	9.92±1.27	45.45	0.09
Rice	7.08±1.30	9.00±1.16	7.85±1.13	10.88	0.9
Bread	7.76±1.28	8.03±1.14	8.29±1.34	6.83	0.9
Fish	8.60±1.25	8.77±1.18	7.02±1.14	–18.37	0.2
Seafood	8.43±1.25	9.03±1.18	6.90±1.14	–18.15	0.1
Eggs	8.99±1.19	7.46±1.14	8.69±1.21	–3.34	0.9
Processed meat	8.82±1.22	8.08±1.18	6.95±1.34	–21.20	0.8
Offal	9.09±1.13	8.27±1.19	6.40±1.21	–29.59	0.1
All red meat	7.65±1.18	8.58±1.14	7.80±1.24	1.96	0.5
All white meat	7.85±1.19	8.03±1.13	8.19±1.19	4.33	0.8
Seed oil	7.70±1.20	7.96±1.22	8.31±1.10	7.92	0.3
Olive oil	8.28±1.16	8.88±1.18	6.76±1.25	–18.36	**0.05**
Butter	6.71±1.13	7.00±1.20	10.52±1.20	56.78↑	0.1
Sugar, honey, jam	6.45±1.20	8.23±1.19	9.31±1.19	44.34	0.06
Cakes, pies, pastries, puddings	8.15±1.26	7.98±1.15	7.95±1.23	–2.45	0.9
Ice cream	8.70±1.13	8.92±1.20	6.06±1.30*	–30.34↓	**0.003**
Wine	7.89±1.29	7.55±1.12	8.68±1.35	10.01	0.1
Beer	8.04±1.23	8.55±1.23	7.55±1.20	–6.09	0.9
Coffee	8.12±1.30	8.24±1.13	7.72±1.20	–4.93	0.5

From Palli *et al.* (2000)

[a] From analysis for covariance model including terms for age, sex, centre, smoking habits (never, ex and current), period of blood drawing and total caloric intake (kcal)

[b] Percentage change from I to III tertile (↑ highest and ↓ lowest variation)

* Dunnett test for multiple comparisons significant at the 0.05 level (first level is the reference category)

Table 2. Adjusted means[a] of relative adduct labelling (DNA adducts per 10^9 normal nucleotides) according to tertiles of estimated daily intake of various nutrients (EPIC Italy, 1993–98)

Nutrient	Tertile of consumption				*p*-value for trend *
	Adjusted mean ± SE			% change[b]	
	I	II	III		
Total protein	9.73±1.60	7.13±1.12	7.33±1.66	–32.74	0.9
Animal protein	9.40±1.34	7.77±1.15	6.74±1.44	–39.47	0.9
Vegetable protein	6.92±1.57	8.25±1.15	8.77±1.50	21.09	0.7
Total fat	9.99±1.44	7.01±1.14	6.96±1.54	–30.33	0.09
Animal fat	9.01±1.41	7.56±1.14	7.58±1.44	–15.87	0.6
Vegetable fat	8.86±1.29	8.75±1.13	6.17±1.32	–30.36	**0.009**
Fatty acids					
Total saturated fatty acids	8.57±1.43	7.92±1.14	7.55±1.51	–11.90	0.9
Oleic acid	9.95±1.35	7.79±1.15	5.87±1.46	–41.01	0.03
Total monounsaturated fatty acids	10.20±1.34	7.67±1.15	6.67±1.47	–34.61	**0.008**
Linoleic acid	8.20±1.48	8.09±1.18	7.82±1.40	–4.63	0.4
Linolenic acid	11.04±1.45	6.86±1.14	6.44±1.42	–41.67↓	**0.01**
Total polyunsaturated fatty acids	8.16±1.46	7.91±1.16	8.02±1.39	–1.72	0.2
Cholesterol	8.25±1.38	7.52±1.14	8.35±1.42	1.21	0.5
Carbohydrates	8.08±1.68	6.76±1.14	9.36±1.63	15.84	0.3
Starch	6.46±1.51	7.15±1.16	10.22±1.48	58.20↑	0.1
Sugar	7.84±1.35	7.32±1.15	8.92±1.33	13.78	0.8
Fibre	8.04±1.43	8.03±1.24	8.01±1.52	–0.37	0.1
Alcohol	7.41±1.30	7.93±1.17	8.66±1.29	16.87	**0.05**
Total calories	8.26±1.78	7.84±1.17	7.91±2.04	–4.24	0.9
Minerals					
Iron	10.31±1.52	7.53±1.16	6.26±1.55	–64.70	0.4
Calcium	7.79±1.37	7.71±1.14	8.62±1.40	9.63	0.9
Sodium	9.35±1.42	6.45±1.13	8.34±1.51	–12.11	0.5
Potassium	10.19±1.43	7.11±1.16	6.89±1.39*	–47.90	**0.004**
Vitamins					
Thiamine	9.74±1.42	7.32±1.14	6.91±1.53	–40.96	0.2
Riboflavin	9.58±1.37	7.24±1.13	7.28±1.43	–31.59	0.4
Niacin	8.35±1.46	8.58±1.13	7.02±1.47	–18.95	0.3
Vitamin B6	8.26±1.42	9.11±1.14	6.65±1.40	–24.21	0.2
Folic acid	8.71±1.36	7.21±1.15	8.19±1.35	–6.35	0.4
Retinol	8.96±1.21	7.42±1.16	7.62±1.23	–17.59	0.2
β-Carotene	9.37±1.22	7.63±1.13	7.15±1.23	–31.05	**0.01**
Vitamin E	8.91±1.46	7.51±1.23	7.73±1.30	–15.27	**0.02**
Vitamin C	9.03±1.24	7.28±1.17	7.79±1.21*	–15.92	**0.01**

From Palli *et al.* (2000)

[a] From analysis for covariance model including terms for age, sex, centre, smoking habits (never, ex and current), period of blood drawing and total caloric intake (kcal)

[b] Percentage change from I to III tertile (↑ highest and ↓ lowest variation)

* Dunnett test for multiple comparisons significant at the 0.05 level (first level is the reference category)

DNA repair capacity. One can speculate that fruit and vegetables, or vitamins, may interfere with DNA repair enzymes, but there is in fact no evidence supporting this hypothesis.

In the light of such observations, a possible explanation for the higher levels of bulky adducts among cancer cases than in controls can be found in the concept of cumulative unrepaired DNA damage. What causes cancer would be the total burden of a genotoxic chemical that remains bound to DNA, after the repair processes. This burden may be higher because DNA repair is impaired, because higher levels of carcinogenic metabolites are present (due to genetic or acquired effects) or because of repeated exposures to the same agent.

We also have to consider the limitations of this model. First, the level of measurement error for bulky adducts is not certain, but seems to be high (coefficient of variation around 20–30%). However, the effect of measurement error is to attenuate a relationship, if error is evenly distributed between the compared groups (Copeland *et al.*, 1977). Table 3 shows an example: in this case, the laboratory error was measured by the intra-class correlation coefficient, which, in turn, was used to correct the observed (attenuated) relative risks so as to obtain a more realistic estimate of the true effects on the risk of cancer. Thus, measurement error is expected to blur existing associations rather than reveal false associations.

A second relevant question concerning Example 1 is whether the effect is attributable to specific agents or to fruits and vegetables as a whole. This question has been addressed by Thompson *et al.* (1999) in an experimental study that aimed to test the hypothesis that increased consumption of vegetables and fruits would reduce levels of markers of oxidative cellular damage. Twenty-eight women participated in a 14-day dietary intervention. The primary end-points assessed were levels of 8-hydroxydeoxyguanosine (8-OHdG) in DNA isolated from peripheral lymphocytes, 8-OHdG excreted in urine, and urinary 8-isoprostane F-2α (8-EPG). Overall, the levels of 8-OHdG in DNA isolated from lymphocytes and in urine and the level of 8-EPG in urine were reduced by the intervention. The reduction in lymphocyte 8-OHdG was greater (32 versus 5%) in individuals with lower average pre-intervention levels of plasma α-carotene than in those with higher levels. The results of this study indicate that consumption of a diet that significantly increased

Table 3. Intraclass correlation coefficients (*r*) for the measurement of estrone by different laboratories and resulting observed relative risks given true relative risks of 1.5, 2.0 and 2.5

Laboratory	*r*	Observed relative risk RR_t=1.5	RR_t=2.0	RR_t=2.5
Laboratory 1	0.12	1.1	1.1	1.1
Laboratory 2				
analysis 1	0.82	1.4	1.8	2.1
analysis 2	0.53	1.2	1.4	1.6
Laboratory 3				
analysis 1	0.57	1.3	1.5	1.7
analysis 2	0.65	1.3	1.6	1.8
Laboratory 4	0.90	1.4	1.9	2.3

RR_t, true relative risk
Observed RR = exp (ln RR_t × *r*)
From Hankinson *et al.* (1994).

vegetable and fruit intake led to significant reductions in markers of oxidative cellular damage to DNA and lipids, in contrast to previous studies that were based on administration of single components of diet.

Thus, we may argue that the choice of the type of intervention can be crucial, and in some circumstances positive results can be obtained more easily with a general category (fruit and vegetable intake) than with a specific component.

Example 2: Modulation by genetic susceptibility and the effect of dose; the case of methylenetetrahydrofolate reductase polymorphism

Common genetic polymorphisms have been reported in the gene encoding methylenetetrahydrofolate reductase (MTHFR), the enzyme that produces 5-methyltetrahydrofolate (5-methyl-THF) required for the conversion of homocysteine to methionine. In individuals with the genotype corresponding to a val/val polymorphism, functional effects include elevation of plasma homocysteine levels and differences in response to folic acid supplementation. The metabolic changes associated with the genotype have been reported to modify the risk for chronic disease (e.g., vascular disease and cancer) and neural tube defects in conjunction with folate deficiency. Folate intake requirements may be different in affected individuals to those of normal or heterozygous individuals. The complex interaction between this common genetic polymorphism of MTHFR and folate intake is the focus of intense investigation (Bailey *et al.*, 1999).

In a study in the United States, an inverse association of this MTHFR gene polymorphism with colorectal cancer was found. The inverse association of methionine intake and positive association of alcohol with colorectal cancer were stronger among val/val individuals. These interactions were not seen for colorectal adenomas (Chen *et al.*, 1998, 1999). In another study, the association between the V (val) allele of the MTHFR gene and ischaemic stroke in an elderly Japanese population was examined. In 256 stroke patients and 325 control subjects, the frequencies of the V allele were 0.45 and 0.32, respectively. The odds ratios and 95% confidence intervals adjusted for the other risk factors were 1.51 (1.02–2.23) for the AV (alanine/valine) genotype and 3.35 (1.94–5.77) for the VV genotype, compared with the AA genotype (Morita *et al.*, 1998).

Moderate elevation of plasma total homocysteine (tHCY) level is a strong and independent risk factor for coronary artery disease. The polymorphism in MTHFR, plasma tHCY and folate using baseline blood levels were examined among 293 Physicians' Health Study participants who developed myocardial infarction during up to eight years of follow-up and 290 control subjects. Compared with those with genotype AA, the relative risk (RR) of myocardial infarction were 1.1 (95% CI, 0.8–1.5) among those with the AV genotype and 0.8 (0.5–1.4) for the VV genotype; none of these RRs was statistically significant. However, those with genotype VV had an increased mean tHCY level (mean ± SEM, 12.6 ± 0.5 nmol/ml), compared with those with genotype AA (10.6± 0.3) ($p < 0.01$). This difference was most marked among men with low folate levels (the lowest quartile distribution of the control subjects): those with genotype VV had tHCY levels of 16.0 ± 1.1 nmol/ml, compared with 12.3 ± 0.6 nmol/ml ($p < 0.001$) for genotype AA. Therefore, the modulating effect of MTHFR seems to be exerted through the tHCY level, especially when folate intake is low (Ma *et al.*, 1996).

Methodological issues

Gene–environment interactions and their potential role in chemoprevention trials: some calculations

To assess the potential role of gene–environment interactions in chemoprevention trials, it is essential to know the penetrance of the genetic trait and its prevalence in the target groups. Table 4 presents an attempt to address this complex issue. Let us imagine we wish to screen high-risk families for a highly penetrant gene (*BRCA1*). In this case, the cumulative risk of breast cancer is approximately 80% in the mutation carriers and the prevalence of the mutations in families is about 50%. Let us hypothetically suppose that tamoxifen reduces the risk by 50%. This means that we have to treat 2.5 family members (carriers of the mutation) to prevent one cancer, i.e., to screen five members to have the same result. However, if we aim at the general population, things change dramatically. Now the cumulative risk is 40%, with an absolute risk reduction of 20%, which means a number needed to treat (NNT) of five women among those carrying the mutations. However, since in the general population only 0.2% are mutation carriers

Table 5. Calculation of the number needed to treat in the case of a screening for a low-penetrance gene (*GSTM1* in smokers), and a high-penetrance gene (*BRCA1*), respectively in the general population or in families

	Smokers		*BRCA1*	
	GSTM1-null	*GSTM1* wild	General population	Families
Relative risk	1.34 (1.21–1.48)[a]	1.0	5	10
Cumulative risk	13%	10%	40% [b]	80%
Risk reduction	50%[c]	50%[c]	50% (tamoxifen)[d]	50%
Cumulative risk after intervention	6.5%	5%	20%	40%
Absolute risk reduction	6.5%	5%	20%	40%
NNT in mutation carriers	15	20	5	2.5
Prevalence	50%	50%	0.2%[e]	50%
NNT in whole target population	30	40	2500	5
NNT in all smokers	35			

NNT, number needed to treat

[a] From Vineis *et al.* (1999); the OR for *GSTM1* in smokers was 1.22 (0.96–1.54)
[b] From Hopper *et al.* (1999)
[c] Theoretical maximum reduction in risk of lung cancer due to chemopreventive agent
[d] Theoretical benefit
[e] Coughlin *et al.* (1999)

(1 in every 500), the number we need to screen is as large as 2500 in order to prevent one cancer, based on the rather optimistic (and theoretical) 50% benefit of tamoxifen treatment.

Considering next a low-penetrance gene, *GSTM1*, we might plan to screen smokers for the *GSTM1* genotype and to address chemoprevention only to them. What would be the advantage? In a meta-analysis (Vineis *et al.*, 1999), the risk of lung cancer associated with the *GSTM1* genotype was 1.34 (it was 1.22 if the meta-analysis was restricted to smokers). Therefore, if the cumulative risk of lung cancer in smokers is 10%, it will be 13% among the null *GSTM1* carriers. Let us suppose again that the chemopreventive intervention has a 50% efficacy. This leads to a cumulative risk of 6.5% among smokers who are *GSTM1*-null, with a number needed to treat (NNT) of 15 (1/6.5%). However, since the carriers of the null genotype are 50% in the population, we need to screen 30 individuals to prevent one cancer. Now we repeat the same calculations with the carriers of the wild genotype, ending up with an NNT of 40. Without screening the population for *GSTM1*, we would have an NNT of 35 (the average of the previous two). Clearly, there is very little advantage in screening for a low-penetrance gene if the NNT just increases from 35 to 40.

Validation of biomarkers: bias and confounding

We have considered several examples of contributions that might come from epidemiology for the selection and validation of biomarkers, and particularly surrogate end-point biomarkers. We suggest that, subject to further validation, bulky DNA adducts could become one such surrogate, particularly in view of the cohort study results mentioned above. Also, mutational spectra in cancer genes and measures of gene–environment interaction are potentially important end-points. However, epidemiological studies are prone to several types of bias (Kleinbaum *et al.*, 1982; Murphy, 1976; Hennekens & Buring, 1987).

Information bias is related to material mistakes in conducting laboratory or other analyses, or in

reporting mutations; for example, a distortion arose from incorrect reporting of the *p53* gene sequence in an early paper, which influenced subsequent reports (Lamb & Crawford, 1986). Another example is the measurement of oxidative damage to DNA: routine phenol-based DNA purification procedures can increase 8-hydroxy-deoxyguanosine levels 20-fold in samples that are exposed to air following removal of the phenol (Wiseman *et al.*, 1995). Such gross contamination would seriously bias an epidemiological study if subsets (batches) of samples from different subgroups in the study population (e.g., exposed vs. unexposed) were to undergo different technical procedures subject to different levels of error.

Selection bias is certainly relevant to molecular epidemiology. Consider hospital-based case–control studies of disease risk in relation to metabolic polymorphisms. We can imagine at least three mechanisms by which selection bias (and, more specifically, Berkson's bias) can occur. First, if a person is hospitalized for a specific reason, but has more than one pathological condition, it is possible that the concurrent disease is also associated with the genetic polymorphism(s) under investigation. Second, patients with a certain allele at the polymorphic locus under investigation can have adverse reactions to drugs and be hospitalized for this reason. Third, induction of an enzyme by treatment can influence the phenotypic indicator of genotype. For example, administration of methotrexate can induce hydroxyfolate-reductase by gene amplification; therefore, if in a case–control study we include cancer patients among the controls, we may have a distorted association between the disease under study and hydroxy-folate-reductase activity.

Specific characteristics of bias and confounding in studies on mutational spectra of cancer genes should also be considered. The size of the biopsies that are selected for investigation provides a clear example of the type of selection bias that can occur in studies on cancer genes. In bladder cancer, for example, it is likely that early-stage tumours are too small to allow the urologist to obtain a biopsy large enough for both research and clinical purposes. However, large biopsies tend to correspond to more advanced cases, which in turn may show a higher proportion of mutations in certain genes (Yaeger *et al.*, 1998).

Detection bias is likely to be a common problem in case–control studies in which the risk factor investigated itself leads to increased diagnostic investigations and thus increases the probability that the disease is identified in that subset of persons.

Detection bias can be considered as a form of information bias, in that the probability of identifying the diseased people is conditional on the clinical information collected, which differs between categories of the risk factor.

In molecular epidemiological studies, this may happen if molecular markers of early disease are prospectively analysed in a cohort. This will lead to easier detection of the eventual clinical disease in those who test positively, even if the marker is not necessarily intermediate in the causal chain leading from exposure to disease. For example, exposure to certain agents, such as formaldehyde, can induce micronuclei in mucosal cells of the oral cavity; these, in turn, may lead to earlier detection of oral cancer through subsequent periodic examination of the workers with positive test results. A similar phenomenon can occur with identification of mutated oncogenes or tumour-suppressor genes in exposed workers, well before the onset of clinical disease.

Publication bias is particularly difficult to characterize and quantify. Publication bias refers to the greater probability that studies with positive findings (e.g., those showing an association between *p53* mutation spectrum and exposure) get published. A way to identify publication bias is to plot the result of each study of a particular phenomenon (expressed, for example, as an odds ratio) against its size. In the absence of publication bias, the plot would be expected to show great variability with small samples and lower variability with large samples, around a central value of the true odds ratio. If publication bias occurs, negative results are not published, particularly if they arise from small investigations, and their results thus do not appear in the plot. For example, in the large database available at the International Agency for Research on Cancer on *p53* mutations (Hainaut *et al.*, 1998), the distribution of the proportion of mutations reported by different studies is skewed: all the studies with a proportion greater than 50% had less than 50 cases, while lower proportions were found in both small and large studies. This

distribution does not necessarily imply that publication bias occurred; it might also suggest that large studies were based on heterogeneous populations, with a variable prevalence of mutations, while small studies refer to small subgroups with specific exposures to carcinogens and a genuinely high proportion of mutations.

Confounding occurs when a third variable creates a spurious association between the exposure at issue and the biomarker measurement. Several variables may act as confounders, for example if they modify the expression of oncogenes or tumour-suppressor genes. One such variable is chemotherapy: for example, cytostatic treatment for leukaemia induces characteristic cytogenetic abnormalities in chromosomes 5 and 7. Confounding arises if, for some reason, therapy is related to the exposure at issue. Disease stage is another potential confounder. Therefore, studies that aim to determine the expression of cancer genes in humans should be restricted to untreated patients or specific stages, or statistical analyses should be stratified according to treatment and stage.

The definition of confounder in molecular epidemiology may be more subtle than in traditional epidemiology. In the study of metabolic polymorphisms in relation to cancer risk, a confounder can be an exogenous exposure which is associated not with another exogenous exposure, but rather with gene expression or enzyme induction (Tajoli & Garte, unpublished). For example, ethanol is an inducer of various metabolic enzymes, and, in turn, excessive intake has been related to colon cancer. The observation of an association between the phenotype for CYP2E1 and colon cancer could be ambiguous, since it could be attributed to a genuine role of this metabolic polymorphism or to confounding by ethanol, which on the one hand would increase the risk of colon cancer and on the other would induce the enzyme (thus creating a spurious association between the two). A somewhat different type of example concerns the observed association of the CYP1A2 polymorphism with variation in the risk of colon cancer; this association is plausible since CYP1A2 is involved in the metabolism of heterocyclic aromatic amines. However, it has also been shown that consumption of cruciferous and other vegetables induces the activity of the CYP1A2 enzyme (Kall *et al.*, 1996), and we know that vegetables reduce the risk of colon cancer (Potter, 1996) (perhaps through their content in antioxidants, acting via pathways unrelated to the CYP1A2 enzyme). Therefore, the association between CYP1A2 and colon cancer can be confounded by dietary habits: specifically, a positive association between the CYP1A2 phenotype and colon cancer may be missed or underestimated because protective factors such as cruciferous vegetables induce CYP1A2 (an example of *negative* confounding). In general, the assessment of inducible enzymes is problematic in case–control studies.

In conclusion, the selection and validation of biomarkers for chemopreventive trials should consider the potential limitations of observational studies.

Interpretation of causal pathways

We can refer to the causal model proposed by a philosopher, Wesley Salmon (1984), to devise a framework for the interpretation of causal pathways in carcinogenesis and in the role played by different biomarkers. Salmon proposes two different models for causality, which are supposed to be complementary. The first model was suggested by Reichenbach many years earlier, and was simply based on probabilistic computation (the 'positive relevance' criterion). The basic idea is that two events (A and B) have a common cause C if we can show (1) that the joint probability of A and B is greater than the product of their separate probabilities, i.e., $p(AB)>p(A)p(B)$; (2) that the introduction of C into the equation entirely or at least partially explains the association between A and B. Condition (1) simply expresses the concept that the probability that two events occur jointly is the product of the individual probabilities if the two events are independent, while it is greater than the product if they are not independent. Condition 2 means that the joint occurrence of A and B can be explained by a common third event: for example, subjects A and B both have an angiosarcoma of the liver, and they were both exposed to vinyl chloride (VC) in the same plant. We can calculate that the probability of such joint events by simple chance is very low (condition 1). In addition, hypothesizing that the common exposure is VC, we observe that all the excess risk (observed/expected cases) is explained by such exposure (condition 2).

However, according to Salmon, this line of reasoning is not sufficient, precisely for the reason we have considered in the case of metabolizing enzymes and dietary habits. In that example, we were not able to disentangle the complex relationships between cruciferous vegetables, the *CYP1A2* metabolic polymorphism and the risk of colon cancer. Salmon suggests that a way to establish whether an event is a genuine cause, in addition to the statistical considerations above, is to include it in a process, and establish whether along the process there is what he calls the *propagation of a 'mark'*. In other words, a genuine causal process is one in which you can follow a 'mark' that propagates over the course of time, precisely because the causal events are able to induce a *structural change* that becomes a part of the effect.

This reasoning is quite relevant to the identification of good intermediate-effect markers for chemoprevention. According to a classical example, yellow fingers are a risk indicator for lung cancer: lung cancer patients have yellow fingers more frequently than population controls, thus fulfilling Reichenbach's definition of causality (positive relevance). However, clearly yellow fingers are just an indirect marker of exposure to tobacco smoke, and not a genuine cause of lung cancer. This is because they do not fulfil Salmon's second criterion of causality, the propagation of a marker along the process. In fact, whereas one can identify *p53* mutations in the lung cancers that are characteristic of tobacco carcinogens, thus following the structural changes left by tobacco into the lung, the same does not hold true for yellow fingers.

In the field of chemoprevention, we propose that certain types of DNA adduct may fulfil such requirements to be used as surrogate end-points. Mutations in cancer genes tend to be late events, and they are too rare to be detected in healthy subjects (in a study of 60 normal subjects in Lyon and Paris, no *p53* mutation in serum was found; P. Hainaut, personal communication), although clearly they may be structural 'marks' for cancer. At the other extreme, protein adducts tend to reflect exposure faithfully, but they are not repaired and are probably weaker predictors of cancer. The advantage of DNA adducts is that apparently they represent the *cumulative unrepaired DNA damage*, and thus they may be used to test several types of chemopreventive agent.

Acknowledgements

This research has been made possible by a grant of the S. Paolo Foundation (Torino) to the ISI Foundation. We are grateful to Anthony B. Miller and Jenny Chang-Claude for helpful comments.

References

Bailey, L.B. & Gregory, J.F., 3rd (1999) Polymorphisms of methylenetetrahydrofolate reductase and other enzymes: metabolic significance, risks and impact on folate requirement. *J. Nutr.*, **129**, 919–922

Berwick, M. & Vineis, P. (2000) Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. *J. Natl Cancer Inst.*, **92**, 874–897

Chen, J., Giovannucci, E., Hankinson, S.E., Ma, J., Willett, W.C., Spiegelman, D., Kelsey, K.T. & Hunter, D.J. (1998) A prospective study of methylenetetrahydrofolate reductase and methionine synthase gene polymorphisms, and risk of colorectal adenoma. *Carcinogenesis*, **19**, 2129–2132

Chen, J., Giovannucci, E.L. & Hunter, D.J. (1999) MTHFR polymorphism, methyl-replete diets and the risk of colorectal carcinoma and adenoma among U.S. men and women: an example of gene-environment interactions in colorectal tumorigenesis. *J. Nutr.*, **129**, 560S–564S

Copeland, K.T., Checkoway, H., Holbrook, R.H. & McMichael, A.J. (1977) Bias due to misclassification in the estimate of relative risk. *Am. J. Epidemiol.*, **105**, 488–495

Coughlin, S.S., Khoury, M.J. & Steinberg, K.K. (1999) BRCA1 and BRCA2 gene mutations and risk of breast cancer. Public health perspectives. *Am. J. Prev. Med.*, **16**, 91–98

Goodman, M.T., Hernandez, S., Wilkens, L.R., Lee, J., Le Marchand, L., Liu, L.Q., Franke, A.A., Kucuk, O. & Hsu, T.C. (1998) Effects of beta-carotene and alpha-tocopherol on bleomycin-induced chromosomal damage. *Cancer Epidemiol. Biomarkers Prev.*, **7**, 113–117

Hainaut, P., Hernandez, T.M., Robison, A., Rodriguez-Tome, P., Flores, T., Hollstein, M., Harris, C.C. & Montesano R. (1998) IARC database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualization tools. *Nucl. Acid Res.*, **26**, 205–213

Hankinson, S.E., Manson, J.E., London, S.J., Willett, W.C. & Speizer, F.E. (1994) Laboratory reproducibility of endogenous hormone levels in postmenopausal women. *Cancer Epidemiol. Biomarkers Prev.*, **3**, 51–56

Hennekens, C.H. & Buring, J.E. (1987) *Epidemiology in Medicine*, Boston, Little, Brown

Hopper, J.L., Southey, M.C., Dite, G.S., Jolley, D.J., Giles, G.G., McCredie, M.R., Easton, D.F. & Venter, D.J. (1999) Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study. *Cancer Epidemiol. Biomarkers Prev.*, **8**, 741–747

Hu, J.J., Roush, G.C., Berwick, M., Dubin, N., Mahabir, S., Chandiramani, M. & Boorstein, R. (1996) Effects of dietary supplementation of alpha-tocopherol on plasma glutathione and DNA repair activities. *Cancer Epidemiol. Biomarkers Prev.*, 5, 263–270

IARC (1986) *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans*. Vol. 38, *Tobacco Smoking*, Lyon, IARC

Kall, M.A., Vang, O. & Clausen, J. (1996) Effects of dietary broccoli on human in vivo drug metabolizing enzymes: evaluation of caffeine, oestrone and chlorzoxazone metabolism. *Carcinogenesis*, **17**, 791–799

Kleinbaum, D., Kupper, L.L. & Morgenstern, H. (1982) *Epidemiologic Research*, Belmont, CA, Lifetime Learning Publ., p. 186

Kucuk, O., Pung, A., Franke, A.A., Custer, L.J., Wilkens, L.R., Le Marchand, L., Higuchi, C.M., Cooney, R.V. & Hsu, T.C. (1995) Correlations between mutagen sensitivity and plasma nutrient levels of healthy individuals. *Cancer Epidemiol. Biomarkers Prev.*, **4**, 212–221

Lamb P. & Crawford L. (1986) Characterization of the human p53 gene. *Mol. Cell Biol.*, **6**, 1379–1385

Li, D., Wang, M., Cheng, L., Spitz, M.R., Hittelmen, W.N. & Wei, Q. (1996) In vitro induction of benzo(a)pyrene diol epoxide-DNA adducts in peripheral lymphocytes as a susceptibility marker for human lung cancer. *Cancer Res.*, **56**, 3638–3641

Ma, J., Stampfer, M.J., Hennekens, C.H., Frosst, P., Selhub, J., Horsford, J., Malinow, M.R., Willett, W.C. & Rozen, R. (1996) Methylenetetrahydrofolate reductase polymorphism, plasma folate, homocysteine, and risk of myocardial infarction in US physicians. *Circulation*, **94**, 2410–2416

Morita, H., Kurihara, H., Tsubaki, S., Sugiyama, T., Hamada, C., Kurihara, Y., Sindo, T., Oh-Hashi, Y., Kitamura, K. & Yazaki, Y. (1998) Methylenetetrahydrofolate reductase gene polymorphism and ischemic stroke in Japanese. *Arterioscler. Thromb. Vasc. Biol.*, **18**, 465–469

Murphy, E.A. (1976) *The Logic of Medicine*, Baltimore, Johns Hopkins University Press

Palli, D., Vineis, P., Russo, A., Berrino, F., Krogh, V., Masala, G., Munnia, A., Panico, S., Taioli, E., Tumino, S., Garte, S. & Peluso, M. (2000) Diet, metabolic polymorphisms and DNA adducts: The EPIC-Italy cross-sectional study *Int. J. Cancer*, **87**, 444–451

Peluso, M., Airoldi, L., Magagnotti, C., Fiorini, L., Munnia, A., Hautefeuille, A., Malaveille, C. & Vineis, P. (2000) White blood cell DNA adducts and fruit and vegetable consumption in bladder cancer. *Carcinogenesis*, **21**, 183–187

Perera, F. (2000) Clues to cancer etiology and prevention from molecular epidemiology. In: *Molecular Epidemiology: A New Tool in Cancer Prevention*. Keystone Symposia, Taos (USA), February 10–15, 2000

Potter, J. (1996) Nutrition and colorectal cancer. *Cancer Causes Control*, **7**, 127–146

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World*, Princeton, NJ, Princeton University Press

Tang, D., Santella, R.M., Blackwood, A.M., Young, T.L., Mayer, J., Jaretzki, A., Grantham, S., Tsai, W.Y. & Perera, F.P. (1995) A molecular epidemiological case-control study of lung cancer. *Cancer Epidemiol. Biomarkers Prev.*, **4**, 341–346

Thompson, H.J., Heimendinger, J., Haegele, A., Sedlacek, S.M., Gillette, C., O'Neill, C., Wolfe, P. & Conry, C. (1999) Effect of increased vegetable and fruit consumption on markers of oxidative cellular damage. *Carcinogenesis*, **20**, 2261–2266

Toniolo, P., Boffetta, P., Shuker, D.E.G., Rothman, N., Hulka, B. & Pearce, N., eds (1997) *Application of Biomarkers in Cancer Epidemiology* (IARC Scientific Publications No. 142), Lyon, IARC

Trizna, Z., Hsu, T.C. & Schantz, S.P. (1992) Protective effects of vitamin E against bleomycin-induced genotoxicity in head and neck cancer patients in vitro. *Anticancer Res.*, **2**, 325–327

Trizna, Z., Schantz, S.P., Lee, J.J., Spitz, M.R., Goepfert, H., Hsu, T.C. & Hong, W.K. (1993) In vitro protective effects of chemopreventive agents against bleomycin-induced genotoxicity in lymphoblastoid cell lines and peripheral blood lymphocytes of head and neck cancer patients. *Cancer Detect. Prev.*, **17**, 575–583

Vineis, P. & Perera, F. (2000) DNA adducts as markers of exposure to carcinogens and risk of cancer. *Int. J. Cancer* (in press)

Vineis, P., Malats, N., Lang, M., d'Errico, A., Caporaso, N., Cuzick, J. & Boffetta, P., eds (1999) *Metabolic Polymorphisms and Susceptibility to Cancer* (IARC Scientific Publications No. 148), Lyon, IARC

Vulimiri, S.V., Wu, X., Baer-Dubowska, W., de Andrade, M., Detry, M., Spitz, M.R. & DiGiovanni, J. (2000) Analysis of aromatic DNA adducts and 7,8-dihydroxy-8-oxo-2′-deoxyguanosine in lymphocyte DNA from a case-control study of lung cancer involving minority populations. *Mol. Carcinogenesis*, **27**, 34–46

Wiseman, H., Kaur, H. & Halliwell, B. (1995) DNA damage and cancer: measurement and mechanism. *Cancer Letters*, **93**, 113–120

Yaeger, T.R., de Vries, S., Jarrard, D.F., Kao, C., Nakada, S.Y., Moon, T.D., Bruskewitz, R., Stadler, W.M., Meisner, L.F., Gilchrist, K.W., Newton, M.A., Waldman, F.M. & Reznikoff, C.A. (1998) Overcoming cellular senescence in human cancer pathogenesis. *Genes Dev.*, **12**, 163–174

Corresponding author:

P. Vineis
University of Torino and CPO-Piemonte,
Via Santena 7,
Torino,
Italy