Defining and Analyzing Cohorts Using Molecular Markers of Cancer Risk

Steven D. Mark*

Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, Bethesda, Maryland

Abstract Cancer is currently regarded to be the phenotypic expression of an accumulation of heritable alterations in the regulators of cell growth and differentiation. Though detailed knowledge of the sequence and in vivo mechanistic effects of these alterations is rudimentary for most, if not all, cancers, their identification does offer the potential for classifying groups of individuals who are heterogeneous with respect to their cancer risks, into more nearly homogeneous subgroups. In this paper, we illustrate the value of using markers, which we define as any manifestation of cellular molecular diversity, to increase subgroup homogeneity. In the context of time-to-event data, we demonstrate for both somatic mutations (acquired p53 abnormalities in gastric mucosal cells) and inherited polymorphisms (polymorphisms in the phase 1 and 2 detoxifying enzymes) how knowledge regarding the population frequency of the marker, the effect of the marker on the risk of cancer development, and/or the effect of the marker on response to therapy, can be used to plan and analyze such trials. Using as paradigms demographic features of the recently begun Shandong precancerous gastric lesion intervention trial, and the recently completed α-tocopherol β-carotene (ATBC) lung cancer prevention study, we review the information, assumptions, and mathematical structure required for planning cancer prevention trials. We graphically demonstrate how informative markers make available strategies for selection, stratification, and optimal weighing, which, when properly implemented, increase the power of tests of effective cancer prevention agents. J. Cell. Biochem. 25S:69–79. © 1997 Wiley-Liss, Inc.†

Key words: biological markers; cancer; gastric cancer; genetics; log rank test; lung cancer; molecular biology; molecular epidemiology; polymorphisms; p53; prevention; power; sample size; survival analyses; trials

Randomized trials are generally the preferred means for testing the effectiveness of a therapeutic intervention. Though no one suggests that randomized trials do not provide equally as convincing an evaluation of cancer prevention agents, trials of preventive therapies occur less frequently than trials of treatment therapies. One reason for this discrepancy is the increased size of the experiment necessary to achieve a powerful test of a preventive intervention. Though we examine subsequently in more detail the determinants of sample size, for now we observe that the number of subjects required for a study and/or the duration of the study generally decreases as the frequency of the outcome under study increases. It is rare to find a cancer in which less than 5% of untreated persons with the disease show signs of progression in a 5-year period. It is rare to find a group of cancer-free individuals who have a 5-year cumulative incidence of a specific cancer as high as 5%. Therefore, even when randomized trials of cancer-prevention agents focus on high-risk groups, large study populations or long study times are required.

The basic considerations that affect cohort design and analysis have been well-established for years [1,2]. What have undergone and continue to undergo dramatic change are the ability to characterize the molecular diversity of cells, the empirical knowledge of the association of this diversity with progression to malignancy, and the increasingly detailed hypotheses one can generate about the steps required for the development of a malignant phenotype. In this paper we refer to measurable manifestations of cellular molecular diversity as markers, and examine how the measurements of such markers, coupled with knowledge about

*Correspondence to: Steven D. Mark, MD, ScD, Division of Cancer Epidemiology and Genetics, Biostatistics Branch, National Cancer Institute, Executive Plaza North, Room 403, 6130 Executive Blvd., MSC 7368, Bethesda, MD 20892-7368.

Received 1 December 1995; Accepted 18 April 1996

their association with cancer, may offer opportunities for increasing the efficiency, and hence extending the feasibility, of cancer prevention trials. Though not required by the mathematical formulations that underlie the results in this paper, for ease of presentation we assume that markers are dichotomous, with an individual either being marker-positive (MP) or marker-negative (MN) at the start of a study.

The potential use of markers in prevention trials is usually divided into two categories: 1) the marker serves as a substitute, or surrogate, for cancer, and trials study the effect of an agent on the marker rather than on the cancer; 2) the marker is used in the design and analysis of studies whose primary endpoint is cancer incidence. Using markers as surrogates has enormous appeal since changes in marker status typically occur with a much higher incidence than changes in cancer status. However, when the question of relevance is the effect of an intervention on cancer and on mortality, and not the effect of the intervention on changes in marker, deducing how a particular intervention's impact on the surrogate translates into its impact on cancer incidence requires a detailed knowledge of the natural history of the progression of the marker to cancer in both untreated and treated individuals. Direct evidence bearing on the latter is only available if the intervention study has already been done and both markers and cancer monitored. Although generating and testing hypotheses about the relations between markers and cancer is an important research activity, the accuracy of deductions about cancer incidence based on changes in marker incidence is unknown.

The focus of this paper is on using markers to aid in the design and analysis of prevention trials whose endpoint is cancer. To illustrate the knowledge required to plan and analyze these studies, we will use the demographic features from two studies. The first example is drawn from a recently initiated prevention study in Shandong, China. This example demonstrates the impact of sub-categorizing the population in terms of somatically acquired differences in marker status. The second example is based on the recently completed α-tocopherol β-carotene (ATBC) study. Features of this study will be used to explore the potential importance of genetically inherited polymorphisms. For both examples, the primary mode of explication will be graphical. Details of the algebraic formulas that underlie these calculations are readily available [1]. The author will gladly supply specifics on how these underling formulations are used in any particular calculation.

QUANTITATIVE CONCEPTS

We consider trials in which individuals are randomly assigned either to a group that receives a single active intervention (AI) or a single placebo (PL). The primary interest is in assessing whether the active intervention prolongs the time a person remains free of a specific cancer. Studies which use as their endpoint the length of time that individuals are free of cancer are referred to as time-to-event (or survival) analyses. In terms of data collection, the principal difference between a time-to-event analysis and the simpler cumulative incidence analysis is that the latter only requires counts of the number of cancers at the end of the study, whereas the former requires information on the actual times at which people develop cancer. Though preference for a time-to-event analysis may be justified by the increase in power afforded by this additional information, in the context of most cancer prevention studies, the major benefit of adopting a survival approach is that it provides a superior framework for considering and displaying the biological considerations that impact on the design and analysis of these studies.

Evaluating the effectiveness of an intervention on the cancer-free time of a population requires comparing the cancer incidence rates of the AI and PL groups. Though our interest is focused entirely on the incidence rates of the specific cancer (CA) under study, we must also be concerned with the disease rate for the competing risks (CR). Competing risks are a heterogeneous category that includes any event which occurs before the end of the study and removes (censors) a cancer-free individual from observation. Though competing risks may arise from a variety of causes, we assume that only deaths from other diseases (including deaths from cancers other than the particular cancer under study) can censor an individual. The rates of the CA and CR diseases are called hazards, and are designated by the symbols $\lambda_{CA}(u)$ and $\lambda_{CR}(u)$, respectively. Figure 1 plots the mortality rates from cancer and from all causes of death for white U.S. males in 1992 [3]. As one would expect, the hazard (incidence) of cancer deaths (and of all deaths) increases with

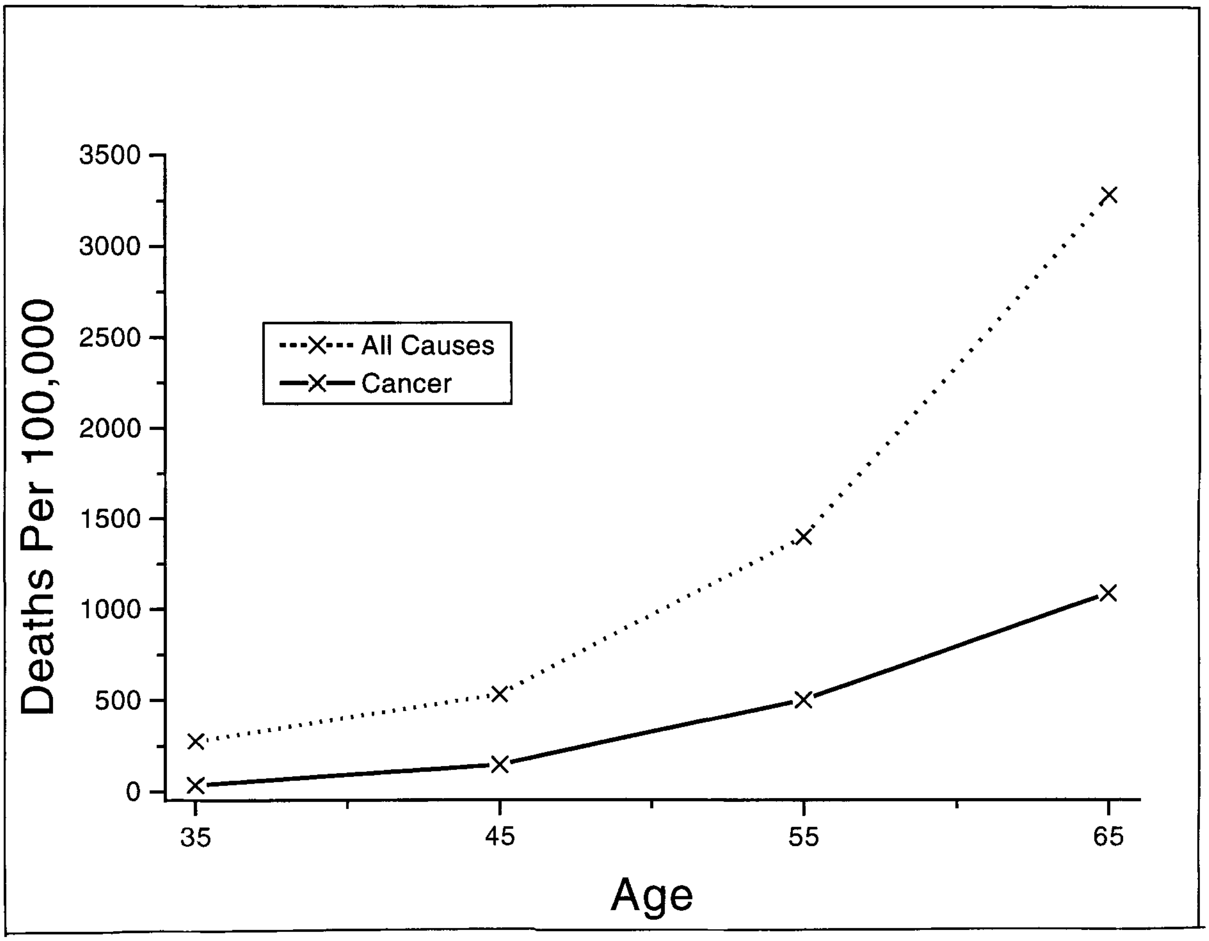

Fig. 1. Mortality rate (hazard of death) versus age in U.S. white males.

age. A simple model, known to be a reasonable fit for cancer hazards in adults [4], is that cancer rates increase as a power of age: λ (age) = b • $(\text{age})^k$. For both cancer (henceforth cancer will always mean the specific cancer of interest) and competing risks, we assume that this formula correctly describes the relation between hazards and age. For our examples, we estimate the b and k for the cancer and for the competing risk from observations made on relevant populations (estimation not shown). To simplify the presentation, we assume that everyone in the study is the same age. The symbol $\lambda_{CA}(u)$ indicates the cancer hazard of this age person u years after the beginning of the study.

If the hazards of the CA and CR disease processes and their interrelation (we make the standard independence assumption [4]) for both the AI and the PL group are known for the entire length—T—of the study, the number of persons in each group who develop cancer, and the number who are alive and cancer-free, is known for all times $u \leq T$. However, even were such elaborate knowledge available, an overall measure of the effect of the intervention on cancer risk would be desirable. One natural measure is to sum the differences in cancer hazards at all times u:

$$L(T) = \int_0^T [\lambda_{CA}(u|AI) - \lambda_{CA}(u|PL)]\, du.$$

L(T) is a negative number if treatment reduces the risk of cancer ($\lambda_{CA}(u|AI) < \lambda_{CA}(u|PL)$); a positive number if treatment increases the risk of cancer ($\lambda_{CA}(u|AI) > \lambda_{CA}(u|PL)$); L(T) is equal to zero if the null hypothesis is true and the chance of death from cancer at all times is identical in both groups ($\lambda_{CA}(u|AI) = \lambda_{CA}(u|PL)$). Unfortunately, at the completion of a randomized trial, the underlying hazards in the AI and PL group are not known, so L(T) is not a practical summary. One intuitive modification of L(T) is the log-rank statistic, LR(T). LR(T) is calculated by summing the difference between O_i—the number of new cancers observed in the AI group at time I—and E_i—the number of the new cancers expected to come from the AI group provided that the null hypothesis were true. If at each of the D times a cancer develops, we keep track of how many cancer-free people are in the AI group, and how many are in the PL group, then

$$LR(T) = \sum_{i=1}^{D} (C_i - E_i)$$

can be computed from the data [1,4]. Like L(T), LR(T) will have its sign and magnitude determined by the direction and extent of differences between $\lambda_{CA}(u|AI)$ and $\lambda_{CA}(u|PL)$.

The usual test of an intervention in a time-to-event study is obtained by dividing the log-rank statistic by an estimate of its variance [1,4]. We call such a test the log-rank test. The magnitude that the log-rank test must achieve for an investigator to conclude that the intervention is either beneficial or detrimental is a function of the size of the type 1 error level (type 1 error levels are described in any introductory statistics book). In our examples, we choose the "usual" two-sided level of .05. The power of an intervention study is the probability that the study will result in a log-rank test negative enough to conclude that the intervention is effective. Examining LR(T), we see that the magnitude of this test will depend on the total number of cancers (D) and, on average, how much smaller O_i is than E_i. This difference depends on the relationship between $\lambda_{CA}(u|AI)$ and $\lambda_{CA}(u|PL)$. Expressing this relationship in terms of a ratio, or relative risk, the smaller the intervention relative risk RR_I(u)

$$RR\_I(u) = \frac{\lambda_{CA}(u|AI)}{\lambda_{CA}(u|PL)},$$

the more negative, on average, is O_i-E_i. Adopting the standard assumption that the intervention relative risk is constant over the time period of the study (RR_I(u) = RR_I for all u), a simple, accurate, and conceptually useful formula that relates the underlying hazards and the intervention relative risk to the required size of the study, N, is

$$N \propto \frac{1}{(\ln RR\_I)^2 \cdot \Pr\,[CA]} \qquad (1)$$

Here Pr[CA] is the probability that a participant will be observed to develop a cancer during the T years of the study. This probability depends upon the underlying rates of cancer and competing risks.

SHANDONG: MARKER INFLUENCES DISEASE RATE BUT NOT DISEASE RESPONSE TO INTERVENTION

The county of Linqu, China (Shandong Province), has a high rate of stomach cancer and a high prevalence of gastric histologic abnormalities [5,6]. These histologic abnormalities have been demonstrated, both in Shandong and elsewhere, to be prognostic for the development of stomach cancer [6,7]. Since 1989, approximately 3,500 members of this population have been participating in a cooperative study between the NCI and the Beijing Institute for Cancer Research. This observational trial involved endoscopic examination and biopsies of gastric mucosa in 1989 and 1994. When classified by their most severe lesion, the prevalence of the abnormalities in 1989 were 46% gastritis (96% of this was chronic atrophic gastritis), 33% intestinal metaplasia, and 21% dysplasia [6]. From preliminary results, we estimate the overall rate of cancer to be 4 per 1,000 person-years, with approximate relative risks of 3.5 and 7 for intestinal metaplasia and dysplasia, respectively. The $\lambda_{CR}(u)$ is estimated to be three times as large as $\lambda_{CA}(u)$.

Though predictive of risk, the histologic categories based on microscopic morphologic examination are thought to result in a classification system which combines biologically heterogeneous groups. In particular, the dysplasias (the large majority of which are mild dysplasias) probably represent a mixture of lesions that differ with regard to both history and destiny. One population of the mild dysplasias likely reverts to normal appearance and function, whereas other populations have acquired irreversible genetic changes. A number of studies using a variety of techniques and markers have examined the frequency of selected genetic abnormalities in cross-sectional collections of gastric cancers and their precursors [8,9]. Abnormalities of the tumor suppressor gene p53 have frequently been found. Though an array of markers would probably provide more information than a single marker, for explication we focus only on p53. To further simplify, no distinction is made with regard to the nature of the p53 abnormality; individuals are classified as marker-negative (MN) if they have no detectable p53 abnormality, and marker-positive (MP) if they have any detectable p53 abnormality. In accord with very preliminary results from a small subset of this population [unpublished results], we assume that the overall frequency of p53 abnormalities is 7.5%, with no MP amongst those with gastritis, 10% MP among those with intestinal metaplasia, and 20% MP among those with dysplasia.

Our goal is to demonstrate conditions under which characterizing individuals by p53 status has a favorable impact on power, sample size, and/or study duration. To this end, we compare a set of hypothetical Shandong intervention studies in which the frequency of MP remains set at 7.5%, but the prognostic importance, or relative risk, of MP varies. During the time course of any given study, this marker relative risk, RR_MP(u) = $\lambda_{CA}(u|MP)/\lambda_{CA}(\mu|MN)$, is assumed to be constant: RR_MP (u)= RR_MP. To find a realistic upper bound for the relative risk, we pretend that the elevation in relative risk associated with dysplasia is entirely due to those 20% who are MP-positive. Under this assumption, an upper bound is RR_MP = 30. For a lower bound we assume the RR_MP = 1: that is, p53 abnormalities convey no risk at all. Throughout this section we assume that p53 has no influence on the effect of the intervention (RR_I).

One option in using p53 in an intervention study would be to screen the population and enroll only people who are p53 positive. We call such a study a Marker Positive Study and contrast it with an All-comers Study in which participants are not screened. The dashed lines (All-comers) and solid lines (Marker Positive) in the graphs of Figure 2 show how, for a power of 90%, the sample size varies with changes in RR_MP. A feature common to both types of studies is the decrease in sample size as RR_MP increases. For both studies, RR_MP increases the probability that a random individual will get cancer. Thus, the decreases in sample size correspond with the observation that more common events require smaller samples. Also anticipated is the increased number of persons required when one goes from a 10-year study (panel A) to a 5-year study (panel B). Notice that, because cancers increase as a power of age, a 5-year study requires that we start with more than twice the number of individuals than a 10-year study. Comparing panel C to panel A shows the smaller sample sizes required by a more effective intervention. Because seemingly small changes in RR_I give rise to large differences in sample size, overestimating the benefit of an intervention can lead to assembling studies with inadequate numbers of participants.

There are two features of the All-comers curve in panel A that could not be anticipated from the sample size formula given in the quantitative section. Though the number of cancers increases as the RR_MP increases, the sample size of the All-comers study begins to increase rather than decrease at high RR_MP (RR_MP 20). The path of the cross-hatched line on Figure 2A also indicates that the All-comers study is not efficiently using the information gathered when RR_MP is large. This line indicates the number of MP individuals participating in a All-comers study at each given level of RR_MP. When RR_MP ≈ 25, the cross-hatched line intersects the solid line: for larger RR_MP, an All-comers study requires more MP-positive persons than a Marker Positive study in order to obtain the same power.

Figure 3 contains a graphical illustration of what underlies these inefficiencies. At the start of the study, randomization assures that the prevalence of MP people is (on average) identical in both treatment groups regardless of the risk of the marker (solid line in Fig. 3A). When the marker conveys a risk (RR_MP ≠ 1) and the treatment has an effect (RR_I ≠ 1), the prevalence of MP in the two treatment groups does not remain equal as the study progresses. Figure 3A shows that, under the conditions of the Shandong example, an excess of MP people accumulate in the AI compared to the PL group. This excess becomes greater as RR_MP increases (both the year-5 and year-10 curves slope upward), and as time increases (year-10 curve is higher than year-5 curve). The accumulation over time of a larger concentration of high-risk (MP) individuals in the AI group causes the relative risk of the cancer incidence (RR_I(u)) to change over time. Thus, despite the fact that within groups defined by marker status the effect of the treatment is constant throughout the study (RR_I(u) = .666). Figure 3B shows that when one ignores marker status, the RR_I(u) in an All-comers study becomes closer to 1 over time. As can be seen by comparing the RR_MP = 20 and RR_MP = 30 curves in panel B, the larger the RR_MP, the more rapid

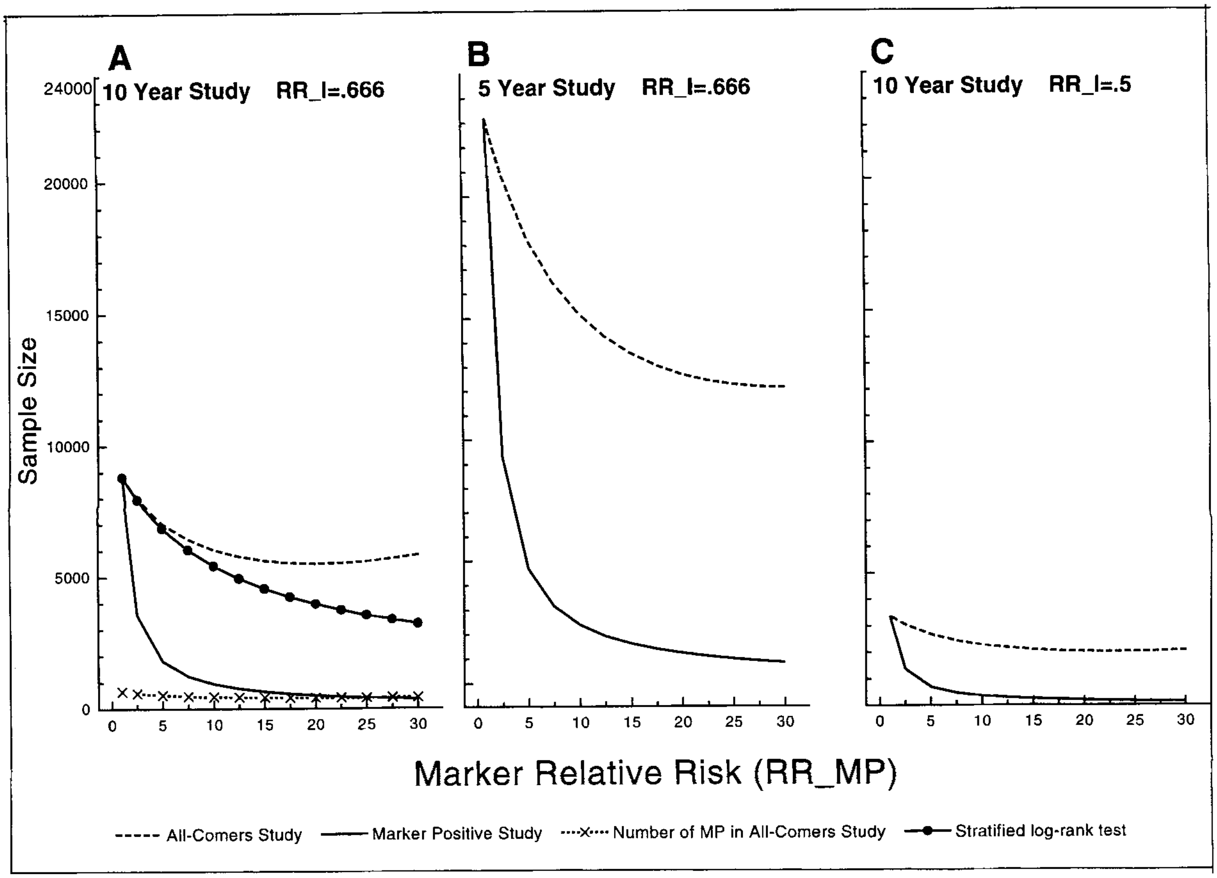

Fig. 2. Sample size requirements in All-comers and Marker Positive studies.

the decrease in the RR_I(u). Graph 2 showed the impact that such a decrease in the magnitude of RR_I can have on sample size.

Figure 4 provides a different illustration of the effect of unmeasured heterogeneity. Suppose we have designed and completed a study for Shandong. Based on our prior experience with this population, we correctly predicted that the 10-year cumulative incidence of gastric cancers would be 11%. If the assumed intervention effect was RR_I = .666, then, in order to achieve 90% power, a sample size of N = 2,253 would have been assembled. Figure 4 graphs the actual power that such a study of 2,253 individuals would have if, instead of arising from a homogeneous population, these cancers arose from a population that consists of two distinct risk groups (7.5% higher-risk MP; 92.5% low-risk group MN), each of which had a RR_I = .666.

We have offered numerous illustrations of the potential impact of heterogeneity, but have proposed only one remedy: screen the population at the outset and study the high-risk individuals. A variety of considerations might render such a plan untenable: there may not be enough high-risk persons in a population; offering a treatment to only one segment of the population may be morally or socially unacceptable; a marker that subdivides a population into more nearly homogeneous subgroups may only become available after the trial is already underway. Fortunately, for the type of heterogeneity we have considered so far, the simple strategy of analyzing the data separately within marker groups, and then combining the results, negates the inefficiencies that arise from analyzing all the subjects together. This analytic strategy, called a stratified log-rank test [2], is implemented by summing the log-rank statistic from each group, summing the variances in each group, and dividing the former by the latter. The line marked with circles in Figure 2A shows the sample sizes required for 90% power in an All-comers study analyzed by the stratified log-rank test. Notice that in contrast to the standard log-rank analysis of the All-comers study (dashed line), the stratified analy-

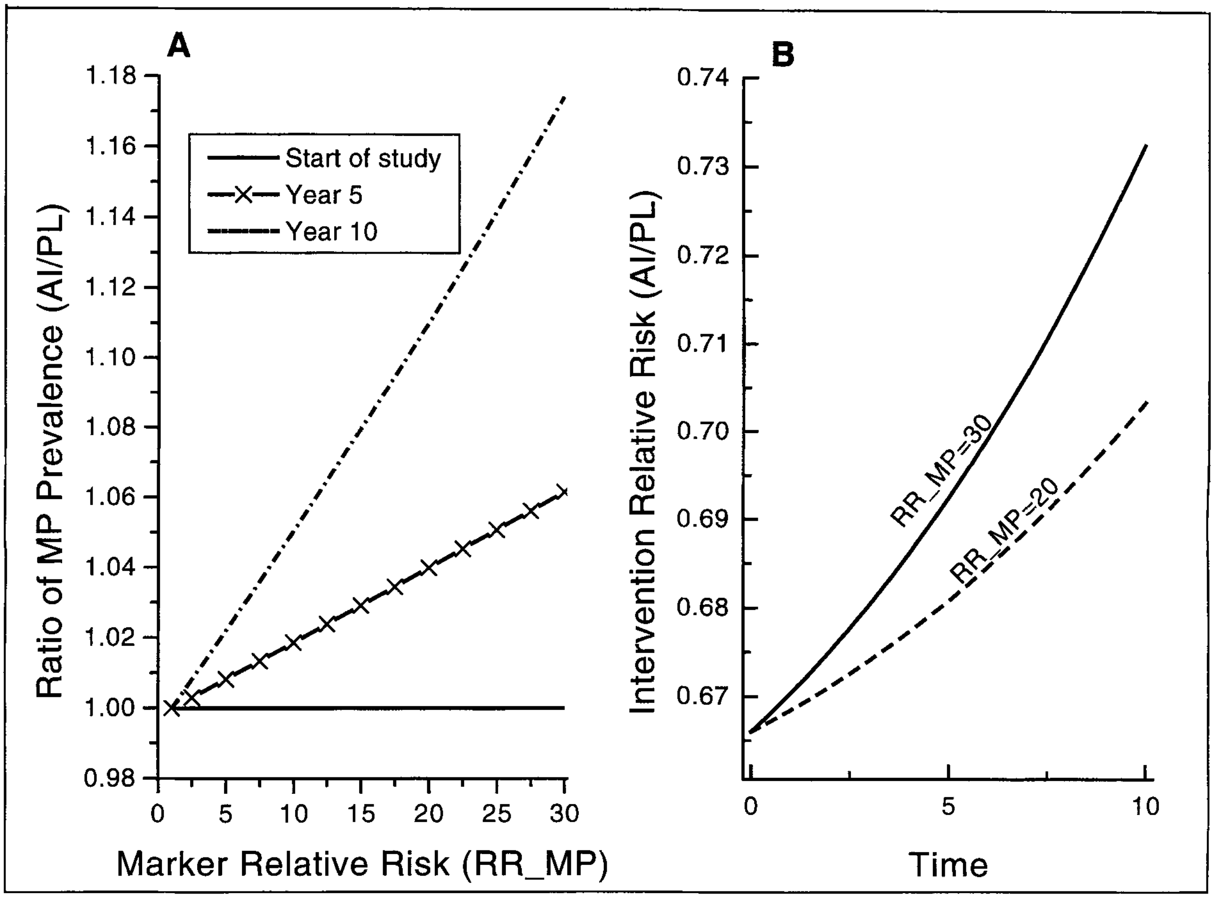

Fig. 3. A: Ratio of MP prevalence versus marker relative risk. B: Intervention relative risk (AI/PL) versus time for 10-year study.

sis no longer has the upturn in sample size requirement when RR_MP > 20. If, as in panel A of Figure 2, the number of MP persons in a All-comers were plotted, this line would always be lower than the sample size line for a Marker Positive study.

THE ATBC LUNG CANCER STUDY: MARKERS INFLUENCE DISEASE RATE AND DISEASE RESPONSE TO INTERVENTION

The Finnish ATBC study [10] provides a framework for examining trials in which marker status affects not only the underlying disease rate, but also the response to treatment. The ATBC study was a prevention trial of 29,133 male smokers, median age 56. Participants were randomized to receive vitamin E, β-carotene, or placebo (2 × 2 design). The study was designed to be of sufficient size to have 85% power to detect a 19% decrease in lung cancer (RR_I = .81). Despite exceeding the predicted number of cancers (876 lung cancers were observed) over the 6 years of follow-up [11], the study failed to find a beneficial effect of either vitamin E or β-carotene on lung cancer. We examine whether, in a study such as this, the power to detect a true beneficial treatment effect of antioxidant therapy might be appreciably diminished by measurable heterogeneity. For simplicity, we assume the intervention consists of only one active antioxidant agent, or one combination of antioxidant agents.

The source of heterogeneity is a germ cell inherited polymorphism that occurs in GSTM1. GSTM1 is a phase 2 detoxification enzyme thought to function in the disposition of several carcinogens, including aromatic hydrocarbons in tobacco smoke [12]. Because of a homozygous deletion [13], approximately 44% of the Finnish populations lacks GSTM1 activity [14]. We designate such persons, often referred to as having the GSTM1 null phenotype, as marker-positive (MP). Though observational studies of lung cancer have differed with respect to the risk con-

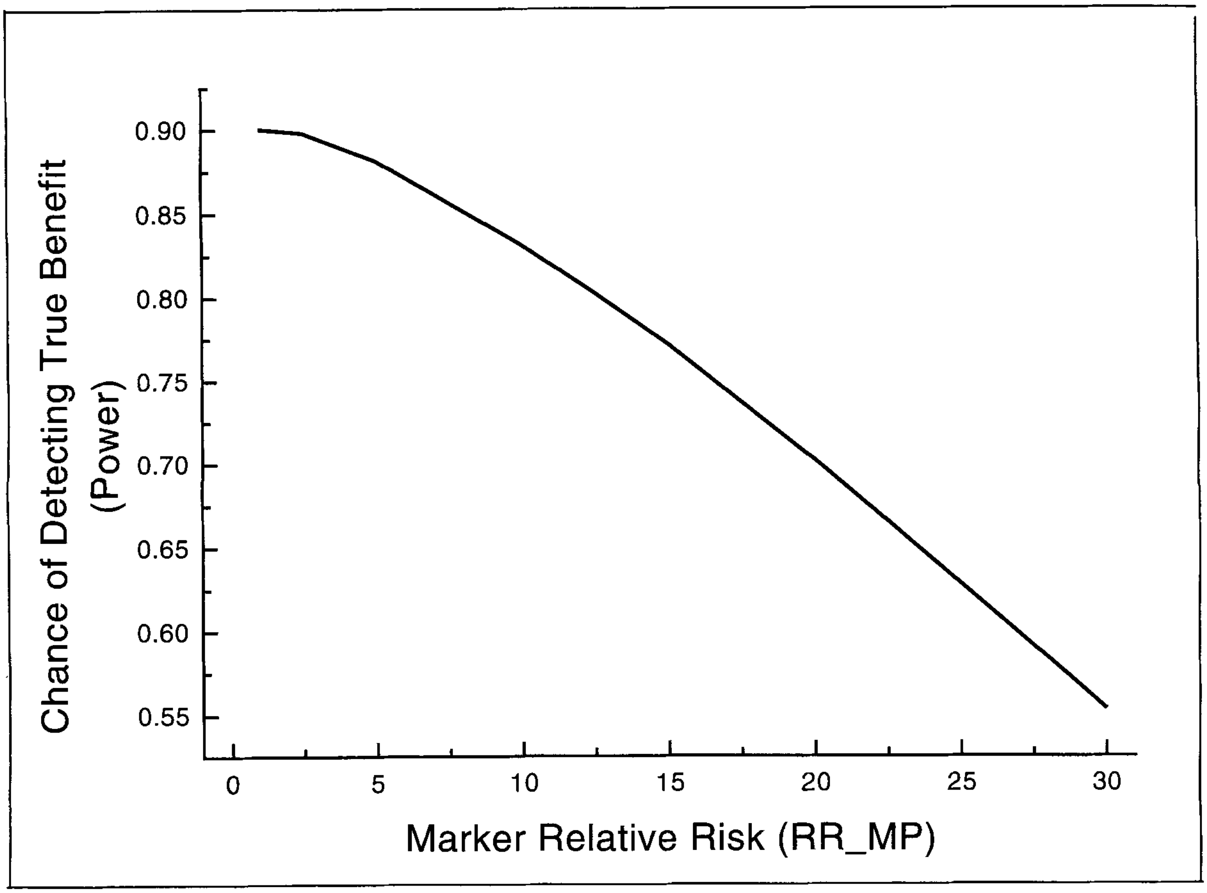

Fig. 4. Power attenuation due to heterogeneity in a cancer prone population.

veyed by the GSTM1 null phenotype, a Finnish case-control study of smokers estimated that GSTM1 null phenotype increases lung cancer risk by a factor of 2 (RR_MP = 2) [14]. This is the estimate of marker risk used in the following examples.

GSTM1 catalyzes the detoxification of electrophilic substances by conjugation with reduced glutathione. Since the primary mechanism of action of the antioxidants is thought to be the reduction of these same substances, the antioxidants can be imagined to function at least as partial "replacement" therapy in MP individuals. Thus, these agents may have greater effect in the MP than the MN individuals. As usual, the effect of an intervention is given in terms of relative risks. We designate the different responses to treatment in the MP and MN groups as RR_I_{MP} and RR_I_{MN}, respectively.

Figure 5 shows the sample sizes required (power = 85%) for each of 3 sets of different intervention relative risks when either the selection of participants varies (MP versus All-comers), or the analysis of the All-comers study varies. All calculations have $\lambda_{CA}(u)$ and $\lambda_{CR}(u)$ chosen so that the All-comers studies duplicate the observed lung cancer incidence and the observed non-lung cancer deaths (the competing risks) of the ATBC study. One extreme is represented in Figure 5, where $RR\_I_{MP} = .5$ and $RR\_I_{MN} = 1$: this corresponds to the antioxidant therapy completely reversing the risk conveyed by the MP phenotype but having no effect on the MN phenotype. Under this scenario, the size of the ATBC study (30,000 persons) would have been more than adequate: only 8,500 people are required for 85% power using a log-rank analysis in an All-comer study. For another extreme, we keep $RR\_I_{MN} = 1$, but set $RR\_I_{MP} = .81$: the level of effectiveness is equal to the minimum thought by the investigators to be of public health importance [11], but the treatment benefit is entirely restricted to the MP persons. Eighty thousand persons would be required for a power of 85%. In the third set of bar graphs (Fig. 5), the total reduction in popu-

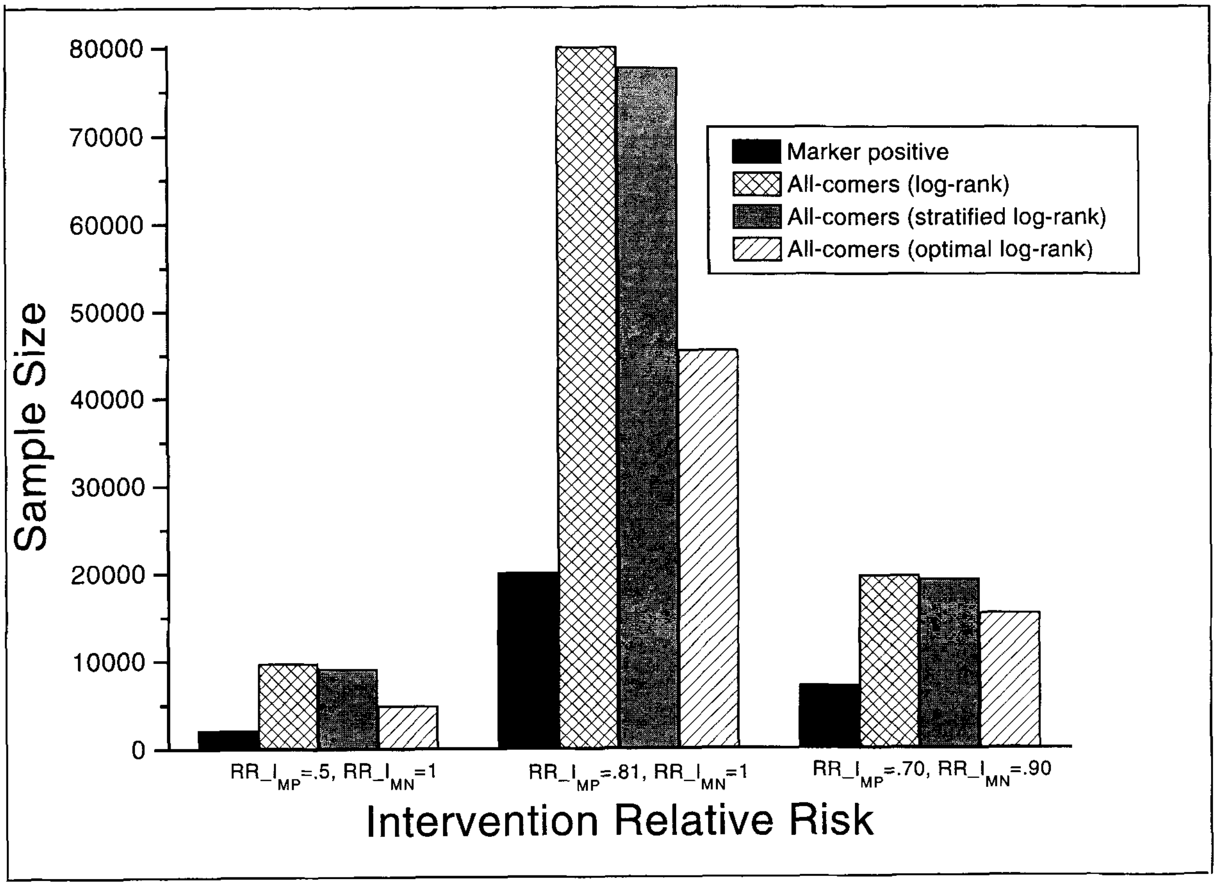

Fig. 5. Variation in sample size when marker affects risk and response to intervention.

lation that lung cancer risk achieved by the intervention is the 19% specified by the investigators, but the benefit is greater in the MP (RR_I$_{MN}$ = .70) than the MN (RR_I$_{MP}$ = .90) individuals.

Notice that the required sample sizes for the MP studies are considerably less than 44% of the required sample size of the All-comers study. Again, this indicates that when one knows marker status, the log-rank test is an inefficient use of the collected information. Here, however, by comparing the appropriate bars in Figure 5, we see that, unlike in the Shandong paradigm, the stratified log-rank aids minimally in the reduction of sample size. When the intervention has no effect on the hazards of the MN people (RR_I$_{MN}$ = 1), the sum of the O_i-E_i in this stratum will be, on the average, zero. However, the variance of this sum will be greater than zero. Thus, if RR_I$_{MN}$ = 1, the contribution of the MN people to the stratified analysis is to add "noise" but no "signal"; the data are more informative if the records of the MN individuals are discarded and the analysis restricted to the MP people. This is exactly what is done by the "optimally weighted" log-rank test. Optimally weighted means that, instead of simply adding together the contributions from each stratum, the contributions are multiplied by a "weight" proportional to the log RR_I in each stratum [1]. This returns us to the "efficient " situation where the size of the Marker Positive study equals 44% of the size of the All-Comers study.

A final example illustrates how marker-derived information on biologic heterogeneity may explain disparate results that arise from identical experiments in two different locations. The frequency of genetic polymorphisms may vary by population. In Japan, though the GSTM1 null phenotype has a prevalence approximately equal to that in Finland, the prevalence of the homozygous m2 CYP1A1 phenotype differs: 11% of Japanese and 2% of Finns are homozygous for this RFLP [15,16]. CYP1A1 is a phase 1 detoxification enzyme. One of its functions is to activate aromatic hydrocarbons

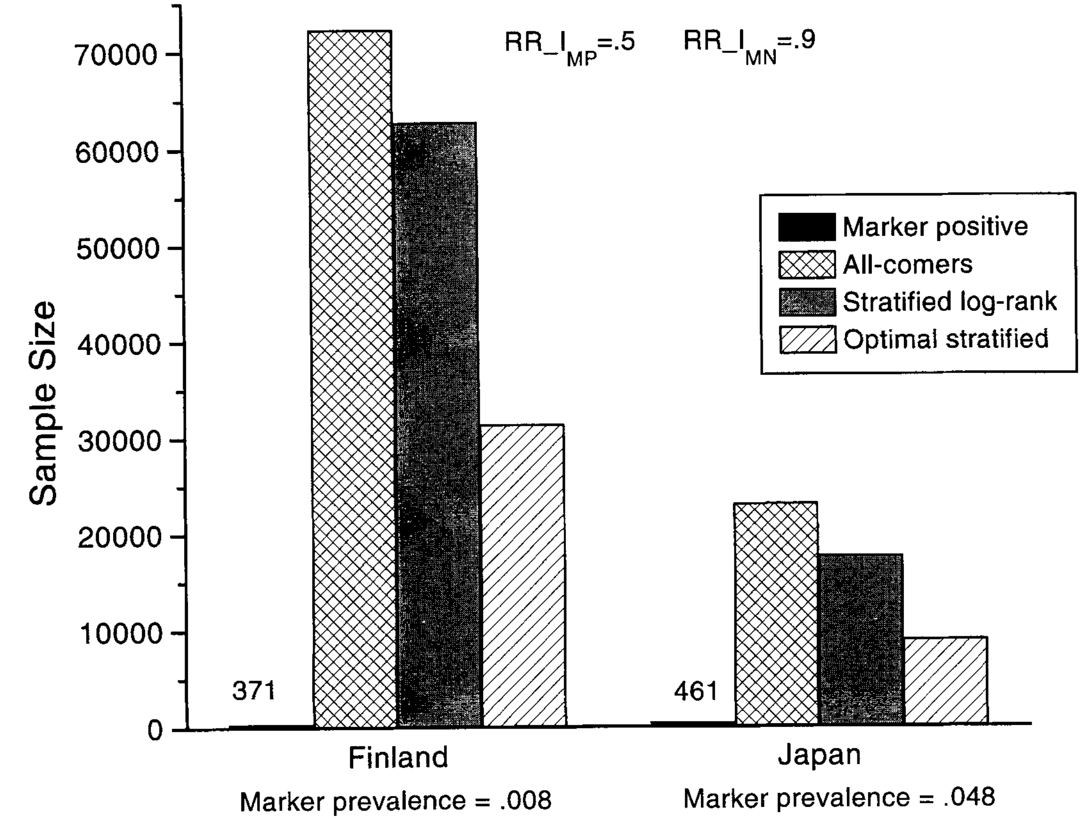

Fig. 6. Effect of variation in population prevalence of marker on study feasibility (RR_I_{MP} = .5, RR_I_{MN} = .90)

to an oxidized state prior to their reduction and conjugation by phase 2 enzymes. Japanese individuals who are both GSTM1 null and CYP1A1 m2 homozygous have been found to have a relative risk of lung cancer of approximately 10 (RR_MP = 10) [15]. Defining the marker-positive state to be the joint presence of both these polymorphisms, and assuming independent segregation, Figure 6 shows the variation in sample sizes between a Finnish and a Japanese study in the presence of identical interventions, identical All-comers cancer incidence, and identical within subgroup intervention effects (RR_I_{MP} = .5; RR_I_{MN} = .9). Equivalently, the point-estimates derived from any analyses of the overall (not stratum-specific) benefit of antioxidant therapy in the two different countries would be discrepant.

DISCUSSION

The accuracy of the standard formulae for calculating sample size requirements, and the efficiency of the usual log-rank test, are predicated upon the assumption that the cancers arise from a population which is homogeneous with respect to both the underlying cancer rate and the magnitude of the intervention effect. Using the known features of two cancer prevention studies, we have constructed examples to illustrate the consequences that occur when the homogeneity assumptions are false. For fixed-marker prevalence, we have shown that the impact on design and analysis are dependent on the biologic repercussions of the heterogeneity. In the Shandong example, the loss of efficiency incurred by employing the usual log-rank analysis rather than the stratified analysis (Fig. 2A), and the attenuation of power based on the standard sample size calculations (Fig. 4), are small for RR_MP <5 and large for RR_MP >10. Similarly, in the ATBC example, the benefit conveyed by using the optimally weighted log-rank test, rather than the usual log-rank test, depends on the extent of the difference between the marker-defined group's response to treatment.

Some important marker-related issues have not been explored in this paper. How should one incorporate uncertainty about marker effect into sample size calculations? Might the measurement of markers of risk influence individuals' compliance with their assigned treatment and affect power in unexpected ways [17,18]? Must "optimal weights" be assigned before the study solely on the basis of a priori beliefs, or can the data be used to estimate weights that are adapted to the nested structure of the log-rank test? If, in the interest of studying groups with high cancer rates, we select marker-positive individuals, might we also be selecting a subset too far along the pathway to cancer to benefit maximally from the intervention? These complicated issues require considerations particular to each population, disease, and intervention under study.

Most cancers are best regarded as the later stage of a disease process rather than as the early stage of a disease. Current evidence suggests that what drives the phenotypic progression toward malignant behavior is the accumulation of genotypic alterations in dividing cells. Though it is likely that some of the key abnormalities responsible for the transformation to cancer have been identified, the information we possess on the sequence, timing, and interaction of these aberrancies, and on the future history of cells at any premalignant stage in the process, is rudimentary. Nonetheless, it is important that randomized cancer prevention trials, as well as observational studies of progression to malignancy, measure and apply the current markers in a manner that is as biologically and statistically coherent as knowledge and knowledgeable hypotheses permit. Hopefully, this will result in immediate gains in terms of more efficient identification of cancer preventive agents. It will probably result in increased information on the dynamics of the progression to cancer. Ultimately, it is increased understanding that will permit more rational design and application of cancer prevention strategies, and better focus and more power in the clinical trials required for their evaluation.

REFERENCES

1. Schoenfeld D (1981): The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 68:316–319.
2. Kalbfleisch JD, Prentice RL (1980): "The Statistical Analysis of Failure Time Data." New York: John Wiley and Sons.
3. National Center for Health Statistics (1995): "Health, United States, 1994." Hyattsville, Maryland: Public Health Service.
4. Cook PJ, Doll R, Fellingham SA (1969): A mathematical model for the age distribution of cancer in man. Int J Cancer 4:93–112.
5. You W-C, Blot WJ, Li J-Y, Chang Y-S, Jin M-L, Kneller R, Zhang L, Han Z-X, Zeng X-R, Liu W-D, Zhao L, Correa P, Fraumeni JF Jr, Xu G-W (1993): Precancerous gastric lesions in a population at high risk of stomach cancer. Cancer Res 53:1317–1321.
6. You WC, Zhao L, Chang YS, Blot WJ, Fraumeni JF Jr (1995): Progression of precancerous gastric lesions [letter]. Lancet 345:866–867.
7. Correa P, Cuello C, Duque E, Burbano LC, Garcia FT, Bolanos O, Brown C, Haenszel W (1976): Gastric cancer in Columbia: III. Natural history of precursor lesions. J Natl Cancer Inst 57:1027–1035.
8. Stemmermann G, Heffelfinger SC, Noffsinger A, Hui YZ, Miller MA, Fenoglio-Preiser CM (1994): The molecular biology of esophageal and gastric cancer and their precursors: Oncogenes, tumor suppressor genes, and growth factors. Hum Pathol 25:968–981.
9. Tahara E (1995): Genetic alterations in human gastrointestinal cancers. The application to molecular diagnosis. Cancer 75 (Suppl):1410–1417.
10. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994): The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med 330:1029–1035.
11. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994): The alpha-tocopherol, beta carotene lung cancer prevention study: Design, methods, participant characteristics, and compliance. Ann Epidemiol 4:1–10.
12. Nebert DW (1991): Role of genetics and drug metabolism in human cancer risk. Mutat Res 247:267–281.
13. Seidegard J, Vorachek WR, Pero RW, Pearson WR (1988): Hereditary differences in the expression of the human glutathione transferase active on trans-stilbene oxide are due to a gene deletion. Proc Natl Acad Sci U S A 85:7293–7297.
14. Hirvonen A, Husgafvel Pursiainen K, Anttila S, Karjalainen A, Pelkonen O, Vainio H (1993): PCR-based CYP*2D6* genotyping for Finnish lung cancer patients. Pharmacogenetics 3:19–27.
15. Nakachi K, Imai K, Hayashi S-I, Watanabe J, Kawajiri K (1991): Genetic susceptibility to squamous cell carcinoma of the lung in relation to cigarette smoking dose. Cancer Res 51:5177–5180.
16. Hirvonen A, Husgafvel-Pursiainen K, Anttila S, Vainio H (1993): The *GSTM1* null genotype as a potential risk modifier for squamous cell carcinoma of the lung. Carcinogenesis 14:1479–1481.
17. Mark SD, Robins JM (1993): A method for the analysis of randomized trials with compliance information: An application to the multiple risk factor intervention trial. Control Clin Trials 14:79–97.
18. Baker SG, Freedman LS (1995): Potential impact of genetic testing on cancer prevention trials, using breast cancer as an example. J Natl Cancer Inst 87:1137–1144.