

The Case-Only Odds Ratio as a Causal Parameter

Paul R. Rosenbaum

Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6302, U.S.A.
email: rosenbaum@stat.wharton.upenn.edu

SUMMARY. In the simplest case-only design, cases of a disease are cross-classified into a 2×2 table describing a genotype attribute and exposure to some environmental agent. In some instances, the genetic attribute has described inherited genes; in other instances, it has described mutations, for instance, damage to proto-oncogenes or tumor suppressor genes leading to cancer. Here, the population case-only odds ratio is written as a causal parameter in terms of potential outcomes with and without exposure to the agent. It is shown that the case-only odds ratio makes sense as a causal parameter with inherited genes, but its magnitude does not have a causal interpretation with mutations, although deviations from 1 do provide information. The difference is that the environmental agent certainly did not cause an individual to inherit particular genes, but it may have caused the mutation.

KEY WORDS: Case-only study; Causal effect; Gene–environment interaction; Observational study.

1. Introduction: Outline, Examples, and Review

The ability to measure genes directly has produced new research designs; see, for instance, the 15 articles in NCI (1999) and the reviews by Khoury and Flanders (1996) and Weinberg and Umbach (2000). One such design is the case-only design for gene–environment interactions (see Piegorsch, Weinberg, and Taylor, 1994; Begg and Zhang, 1994; Botto and Khoury, 2001). In a case-only design, information is obtained only from cases of a particular disease, with no information obtained about people without the disease. In its simplest form, two genotypes are distinguished, and cases are cross-classified into a 2×2 table indicating genotype and exposure to some environmental agent. The odds ratio in this table is taken as a measure of “gene–environment” interaction.

The case-only design has been applied to both inherited genes and to mutations leading to cancer, and Sections 1.1 and 1.2 discuss examples. Section 2 shows the case-only odds ratio is a causal parameter with inherited genes, while Section 3 shows its magnitude has no clear interpretation for mutations. However, with mutations, deviation from an odds ratio of 1 does provide information.

1.1 Example: Smoking, the XPD Gene, and Bladder Cancer

The protein coded by the xeroderma pigmentosum group D (XPD) gene performs several functions, including DNA repair. Specifically, XPD helps to open the DNA helix to permit the removal of a piece of DNA containing a damaged base (Lehmann, 2001). In human populations, there are slight inherited variations (or polymorphisms) in the XPD gene that occur naturally. In particular, at codon 751, either the amino acid lysine (Lys) or glutamine (Gln) may be coded. Polymorphisms in XPD might strongly interact with certain carcinogens: The carcinogens might be more potent if DNA repair is less effective.

Stern et al. (2002) hypothesized that polymorphisms at codon 751 interact with cigarette smoking as a cause of bladder cancer, anticipating from other studies that being homozygous for Gln at 751, Gln/Gln, might reduce risk from smoking. Their study included both case-control analyses and case-only analyses (Table 1). In Table 1, the odds that a case with genotype Lys/Lys or Lys/Gln will be a smoker rather than a nonsmoker are $171/29 = 5.9:1$, whereas the odds that a case with genotype Gln/Gln will be a smoker are $21/9 = 2.3:1$, so Lys/Lys or Lys/Gln cases are $(9 \times 171)/(29 \times 21) = 2.5$ times more likely to be smokers than Gln/Gln cases. Fisher’s exact test yields a one-sided significance level of 0.037.

1.2 Example: Coffee, K-ras Mutations, and Pancreatic Cancer

Table 2 is from a case-only study by Porta et al. (1999) concerning the roles of coffee and acquired K-ras mutations in cancer of the pancreas. Mutations of the K-ras gene are typically found in pancreatic cancers (Amoguera et al., 1988). Table 2 is the basis for one of their several analyses: It describes 107 cases of pancreatic cancer, cross-classified by an environmental agent, namely regular consumption of coffee, and mutation of the K-ras proto-oncogene at codon 12. The ras proto-oncogenes, including K-ras, act on the inner side of the cell membrane and convey growth signals from the membrane (McKinnel et al., 1998, p. 136). The odds ratio in Table 2 is $(73 \times 8)/(16 \times 10) = 3.65$, which differs significantly from 1 by Fisher’s exact test, with one-sided significance level 0.020. Among cases of pancreatic cancer, coffee drinkers were more likely than other cases to have tumors with K-ras mutations.

In thinking about Table 2, it should be kept in mind that prospective studies of pancreatic cancer typically find no association with coffee consumption and a moderately strong association with quantity and duration of cigarette smoking (e.g.,

Table 1*XPD polymorphisms and smoking among bladder cancer cases*

XPD codon 751	Nonsmoker	Smoker	Total
Lys/Lys or Lys/Gln	29	171	200
Gln/Gln	9	21	30
Total	38	192	230

Source: Stern et al. (2002).

Stolzenberg-Solomon et al., 2002; Potter, 2002). The conventional view is that perhaps a quarter of pancreatic cancers are caused by smoking.

1.3 Review: Causal Effects as Comparisons of Potential Responses

The case-only odds ratio will be written as a causal parameter, and this section briefly reviews a notation for causal effects. The case-only odds ratio is typically computed in an observational study of the effects caused by an environmental agent, in relation to genes and health outcomes, that is, a study of the effects of an agent that was not subject to randomized controlled experimentation. In his review of observational studies, Cochran (1965, p. 236) offered the following advice. Attributing the idea to Dorn (1953), Cochran argued that thinking about an observational study should begin by thinking first about an analogous randomized experiment, in which subjects are randomly assigned to either control or exposure to the agent; then, continue by thinking about the consequences for this study of the absence of randomization. The advantage, of course, is that there is a fully adequate and successful theory of causal inference in randomized experiments, pioneered by Fisher (1935), and the first step of this two-step process then builds upon a solid foundation. If the analogous experiment makes sense, then that is a good beginning, and one can go on to address the substantial problems due to the absence of randomization. Alternatively, if the analogous experiment makes little sense, then there is little or no hope that situation will be improved by the absence of randomization, that is, by confounding from observed and unobserved covariates. In a case-only study, information is obtained only about cases; however, these cases might have been the cases of disease generated by an analogous randomized experiment, with the cases retained for analysis and the others, the noncases, set aside. Does the magnitude of the case-only odds ratio have a causal interpretation in a simplified world in which exposures to the agent strike heterogeneous individuals at random?

In the theory of experimental design, to say that exposure to an agent *caused* a particular individual to have a particular

Table 2*K-ras mutations and regular coffee consumption among cases of cancer of the pancreas*

Regular coffee drinker	Yes	No	Total
K-ras mutated	73	10	83
K-ras wild type	16	8	24
Total	89	18	107

Source: Porta et al. (1999).

outcome is to compare the outcome this individual would have had if exposed to the agent with the alternative outcome this individual would have had if not exposed to the agent. That is, each person i has two potential responses, (r_{Ti}, r_{Ci}) , where r_{Ti} is the outcome observed from i if exposed and r_{Ci} is the outcome observed from i if not exposed, and the causal effect of exposure on i is a comparison of r_{Ti} and r_{Ci} (see Neyman, 1923; Rubin, 1974). For instance, if a 1 or a 0 response indicates that a specific disease did or did not occur in a specific time interval, then an individual i with $(r_{Ti}, r_{Ci}) = (1, 0)$ would be diseased if exposed to the agent and not diseased if unexposed, so exposure causes this individual's disease. In fact, one observes from i either r_{Ti} if i is exposed or r_{Ci} if i is not exposed, but one never observes (r_{Ti}, r_{Ci}) jointly, so the causal effect on one person i cannot be calculated. Nonetheless, inferences about causal effects for populations are possible, for instance, in randomized experiments (Fisher, 1935).

If the outcome is binary, say occurrence of a particular disease or not, and if the agent might cause this disease but does not prevent it, then the distribution of possible outcomes may be summarized concisely with a notation similar to that used by Hamilton (1979) for 2×2 tables. There are τ individuals in the analogous randomized experiment, who divide into three groups, $\tau = \beta + \delta + \nu$. Specifically, there are β people i with $(r_{Ti}, r_{Ci}) = (1, 1)$ who would develop the disease whether or not they are exposed to the agent, δ other people with $(r_{Ti}, r_{Ci}) = (1, 0)$ who would develop the disease only if exposed, and ν other people with $(r_{Ti}, r_{Ci}) = (0, 0)$ who would not develop the disease even if exposed. The quantities (β, δ, ν) describe the unobservable joint distribution of quantities (r_{Ti}, r_{Ci}) whose marginal distributions are separately observable, at least in randomized experiments. Typically, this population is heterogeneous in other ways that are neither understood nor recorded.

Different ways of developing the statistical theory of experimental design view the τ individuals slightly differently, but this without consequence for the current discussion. In particular, Fisher (1935) viewed the τ individuals as a finite population, with randomness entering only through the random assignment of treatments, so that randomization forms the "reasoned basis for inference" in randomized experiments. In contrast, Neyman (1935) viewed the responses of the τ individuals as generated by a stochastic model. In the current context, Little and Wright (2003) develop one stochastic model for carcinogenesis with genome instability and clonal evolution, and the heterogeneous population of people under study might have been generated by stopping that model at a moment in time, perhaps different times for different individuals, and then exposing some individuals to the environmental agent. In this article, only population parameters are discussed, not inference from sample to population, so these slight differences are without consequence for the results.

In this population, the risk of disease without exposure to the agent is β/τ , which increases to $(\beta + \delta)/\tau$ with exposure, so the relative risk caused by exposure in this population is $\{(\beta + \delta)/\tau\}/(\beta/\tau) = (\beta + \delta)/\beta$; see Hamilton (1979) for discussion of this effect measure and many others, including various definitions of attributable risk. If a fraction θ of the population were selected at random and exposed to the agent,

Table 3

Exposure effects by genotype: frequencies of potential responses in the population

Case if exposed	Yes	Yes	No
Case if not exposed	Yes	No	No
Genotype A	β_A	δ_A	ν_A
Genotype B	β_B	δ_B	ν_B

with the remaining $1 - \theta$ left unexposed, then the ratio of the proportions of diseased subjects in exposed and unexposed groups is a consistent estimate (as $\tau \rightarrow \infty$) of the relative risk, $(\beta + \delta)/\beta$. See Rosenbaum (2001) for exact inference in this setting and many others with $r_{Ti} \geq r_{Ci}$.

2. Inherited Gene and Environmental Agent

Table 3 describes two genotypes, A and B, and the distribution of responses they would exhibit with and without exposure to the environmental agent. It is analogous to two copies of the situation in Section 1.3, one for A and one for B. Here, there are $\beta_A + \delta_A + \nu_A$ type A's in the population. Of these, β_A would be cases of disease whether exposed or not, δ_A would be cases only if exposed, and ν_A would not be cases even if exposed. As in Section 1.3, for type A's, the relative risk is $(\beta_A + \delta_A)/\beta_A$. The situation with type B's is parallel.

The assumption that commonly underlies a case-only study is that exposures to the agent strike individuals at random, that is, with constant probability θ for all individuals in the population in Table 3 (Piegorsch et al., 1994). This will be called the *independence assumption*. Albert et al. (2001) suggest caution about this assumption, noting ways it can be false, and when false, may distort estimates. The independence assumption can be weakened somewhat by dividing the population into strata, for instance by ethnicity or race; then the exposure rate must only be constant within strata, but may vary between strata. The independence assumption embodies the assumption, mentioned in Section 1.3, that exposure strikes individuals at random with probability θ , but importantly emphasizes or adds that the same probability θ applies for both genotypes, A and B.

Table 4 describes the expected counts for a case-only study derived from Table 3 under the independence assumption. Table 4 contains the cases from Table 3 cross-classified by exposure which occurs at rate θ . The β_A type A's who will be cases whether exposed or not all appear in Table 4, with $\theta\beta_A$ exposed, $(1 - \theta)\beta_A$ not exposed. In contrast, the δ_A type A's who are cases only if exposed appear in Table 3 only if they are exposed, so only $\theta\delta_A$ of them appear in Table 4. And so on.

Table 4

Exposure effects by genotype: expected counts for a case-only study under the independence assumption

	Exposed	Not exposed
Genotype A	$\theta(\beta_A + \delta_A)$	$(1 - \theta)\beta_A$
Genotype B	$\theta(\beta_B + \delta_B)$	$(1 - \theta)\beta_B$

The case-only odds ratio from expected counts is the odds ratio in Table 4, namely:

$$\frac{(\beta_A + \delta_A)\beta_B}{(\beta_B + \delta_B)\beta_A} = \left(\frac{\beta_A + \delta_A}{\beta_A} \right) \left(\frac{\beta_B + \delta_B}{\beta_B} \right)^{-1}, \quad (1)$$

which is the ratio of the relative risk $(\beta_A + \delta_A)/\beta_A$ among type A's and the relative risk $(\beta_B + \delta_B)/\beta_B$ among B's. Under the independence assumption, the case-only odds ratio shows how the effects of exposure are different for the two genotypes, providing effect is measured by the relative risk. This risk ratio equality (1) is essentially the same as the very nice result for odds ratios of Piegorsch et al. (1994, p. 159), but their rare-disease assumption is not needed (see Schmidt and Schaid, 1999). Piegorsch et al. (1994) clearly intended their method to be applied with inherited genes; indeed, their paper begins: "Inherited genetic susceptibility is an important determinant of disease risk." The situation is different with mutations, discussed in Section 3.

3. Mutation and Environmental Agent

3.1 Studies of Tumor Cells

In a study of inherited genes, nearly every cell of the person contains the inherited gene. In contrast, in a case-only study of mutations leading to cancer, the study examines tumor cells. It is important to pause and recall what is involved.

The cell divisions or clonal evolution from a single normal cell to a tumor has been represented as a binary tree by Nowell (1976, Figure 1). A mutation occurs in a cell, and if the cell survives, it may pass the mutation to its descendants. A mutation might affect the development of cancer if it disables a proto-oncogene, such as one of the ras genes. With ras disabled, the cell and its descendants may stimulate themselves to reproduce, without waiting for growth signals from outside the cell, that is, the cell may have developed the "acquired capability [of] self-sufficiency in growth signals" (Hanahan and Weinberg, 2000, p. 58). Such a cell may not gradually produce a small colony of descendants, but instead rapidly produce a much larger colony of descendants, whose many cells are now a larger target for subsequent mutations. Perhaps many years later, one of these (perhaps, by now, quite numerous) descendants suffers mutations that disable the function of the p53 gene. The p53 tumor suppressor gene is involved in regulating both DNA repair and apoptosis, so this cell may have acquired an "enabling capability [of] evading apoptosis" (Hanahan and Weinberg, 2000, p. 61). This one cell and its descendants are reproducing more rapidly, repairing DNA less effectively, and surviving subsequent mutations more often than do normal cells. Quite often, one descendant suffers the loss of a chromosome (aneuploidy), resulting in greater instability of the cell's genome (Lengauer, Kingler, and Vogelstein, 1998). Eventually, the colony of descendants is a clinically noticeable tumor. In patients, the individual mutations in single cells that produced the tumor are not observable; however, many of the cells of the tumor are descendants of that one initial cell, and their genomes provide incomplete information about the sequence of genetic transformations that created the tumor.

If a person develops a specific tumor, say cancer of the pancreas, tumor cells may be examined to determine whether a

particular mutation is present or not in many of the tumor cells, say a mutation in the K-ras gene. If the K-ras mutation is present in tumor cells, we infer that sometime in the past, a common ancestor of these tumor cells suffered this mutation, although that specific ancestor is long gone. This is what is done in a case-only study. If the person does not have the specific tumor—that is, if the person is not a case—then (i) there are no tumor cells to examine and (ii) there is no corresponding place to look for the mutation. At the risk of belaboring this obvious point: If a person does not have an identifiable tumor, then one cannot reasonably examine a sample of healthy tissue—it is extremely unlikely that healthy tissue would have measurable numbers of cells with any specific mutation—and one cannot (practically) examine every single cell in the hope of determining whether the person has some one cell with a K-ras mutation. One can only examine a noticeable tumor much of whose mass shares a common ancestry.

Consider, again, the very simplest situation, an environmental agent that exists in just two versions, exposed at a single, specified moment or never exposed. A single fixed dose of radiation is a traditional example. (Obviously, the “dose” of coffee consumption is more complex; there are doses: cups per day; schedules of doses: began drinking coffee at age 22, quit at age 30, resumed at age 32, etc; methods of preparation: filtered, espresso, etc; however, multiple versions of the agent introduce complexity without introducing new relevant concepts.) With two versions, each person has two potential genetic histories, one if exposed to the agent, the other if spared exposure. Both histories may contain a tumor of the clinical type under investigation, possibly different tumors having origins in different cells, or only one history may contain a tumor, or neither may contain a tumor. As in Section 2, in Hamilton (1979) and Rosenbaum (2001), to keep things simple, it will be assumed that the agent may sometimes cause, but never prevents, cancers or mutations. This is the simplest situation. If the case-only odds ratio with mutations lacks a clear causal interpretation in this simplest situation, then it cannot be relied upon in general.

In principle, one could imagine a randomized experiment. The population is divided at random: with probability θ a person is exposed to the agent, with probability $1 - \theta$ a person is not exposed; these exposures strike some people and spare others completely at random. If exposed, a person has one genetic history, and if spared exposure, has the other. In this imagined experiment, this person would enter our case-only study if the person’s realized history contained a tumor of

the type under study. That tumor would then be examined to determine whether many of its cells have the type of mutation under study.

3.2 Cross-Classification of Potential Responses

Table 5 describes the distribution of potential responses in the population for a study relating a type of tumor, a specific mutation, and an environmental exposure. If there is such a tumor, one can check to see whether many of its cells contain the mutation. In particular, $\beta_+ = \beta_{yy} + \beta_{yn} + \beta_{nn}$ people would develop this clinical type of cancer whether exposed to the agent or not. These β_+ people have two, possibly very different, potential tumors, which may originate in different cells. For instance, a heavy smoker who worked with asbestos might develop a very different lung tumor than he would have developed had he not worked with asbestos, but he might have developed a lung tumor whether or not he worked with asbestos. Table 5 divides these β_+ people into three groups based on whether their tumors would exhibit the mutation under study with or without exposure to the agent. In Table 5, there are: β_{yy} people who would have both the cancer and the mutation whether exposed to the agent or not; β_{yn} other people who would have the cancer whether exposed or not, but would have the mutation only if exposed; and β_{nn} people who would have the cancer whether exposed or not, but would not have the mutation whether exposed or not.

The situation is somewhat simpler for the $\delta_+ = \delta_y + \delta_n$ people who would develop the type of tumor under study only if exposed to the agent. Because these δ_+ people would not have a tumor without exposure to the agent, one can only speak of the presence or absence of the mutation in their tumor when they are exposed to the agent; otherwise, there is no tumor to talk about. There are δ_y people who develop a tumor containing the mutation when exposed to the agent, and δ_n other people who develop a tumor without the mutation when exposed, but these $\delta_y + \delta_n$ people do not develop a tumor without exposure. In addition, ν people do not develop the tumor under study whether exposed to the agent or not, and so, of course, one cannot speak about the presence or absence of the mutation in their tumor. As before, the total number of people is denoted $\tau = \beta_{yy} + \beta_{yn} + \beta_{nn} + \delta_y + \delta_n + \nu$.

3.3 A Randomized Experiment

Following Cochran’s (1965) advice, mentioned in Section 1.3, imagine a randomized experiment in which a single biased

Table 5
Exposure effects for mutation and disease: frequencies of potential responses in the population

Case if exposed	Case if not exposed	Mutation found in tumor if exposed	Mutation found in tumor if not exposed	Frequency
Yes	Yes	Yes	Yes	β_{yy}
Yes	Yes	Yes	No	β_{yn}
Yes	Yes	No	No	β_{nn}
Yes	No	Yes	–	δ_y
Yes	No	No	–	δ_n
No	No	–	–	ν
			Total	τ

Table 6

Mutation and disease in an imagined randomized experiment

Cancer	Mutation found	Exposed	Not exposed
	in tumor		
Yes	Yes	$\theta(\beta_{yy} + \beta_{yn} + \delta_y)$	$(1 - \theta)\beta_{yy}$
Yes	No	$\theta(\beta_{nn} + \delta_n)$	$(1 - \theta)(\beta_{nn} + \beta_{yn})$
No	–	$\theta\nu$	$(1 - \theta)(\nu + \delta_y + \delta_n)$
	Total	$\theta\tau$	$(1 - \theta)\tau$

coin is flipped independently for each person in Table 5, so each person is either exposed to the agent with probability θ or spared exposure with probability $1 - \theta$, where $0 < \theta < 1$. The experiment provides information about cancer cases and about individuals without cancer, while the case-only study provides information only about cases.

The expected distribution of observed responses in this experiment is given in Table 6 which is derived from Table 5. It is important to notice that if, somehow, we did not know the rate of exposure, θ , we could estimate it from the marginal total number exposed, $\theta\tau$, and the number not exposed, $(1 - \theta)\tau$, as $\theta = \theta\tau / \{\theta\tau + (1 - \theta)\tau\}$. Armed with θ or an estimate of it, the information in Table 6 would permit estimation of many interesting causal parameters, including: (i) the number of people who, if exposed, would have a cancer caused by exposure, $\delta_y + \delta_n$, (ii) the number of people who, if exposed, would have a tumor in which the mutation is present and the mutation was caused by the exposure, $\beta_{yn} + \delta_y$, as well as many other causal parameters. Note, however, that the observable, marginal distributions in Table 6 do not fully identify the joint distribution in Table 5; for example, replacing $(\beta_{yn}, \beta_{nn}, \delta_y, \delta_n)$ by $(\beta_{yn} + \epsilon, \beta_{nn} - \epsilon, \delta_y - \epsilon, \delta_n + \epsilon)$ in Table 5 leaves Table 6 unchanged. If the randomized experiment were feasible, it would provide a great deal of useful information about the effects of the exposure in causing cancers and mutations. How does this experiment compare to the case-only study?

3.4 *The Case-Only Odds Ratio with Mutations*

Imagine that the experiment in Section 3.3 had been performed, but only data from cases was recorded. Table 7 is the portion of Table 6 describing the cases.

It is important to note that if we do not know the rate, θ , of exposure to the agent, then we *cannot* estimate it from Table 7. Unlike Table 6, the ratio of the total number exposed to the total number not exposed in Table 7, is *not* $\theta / (1 - \theta)$.

Table 7

Exposure effects for mutation and disease: expected frequencies for a case-only study

	Exposed	Not exposed
Mutation	$\theta(\beta_{yy} + \beta_{yn} + \delta_y)$	$(1 - \theta)\beta_{yy}$
No mutation	$\theta(\beta_{nn} + \delta_n)$	$(1 - \theta)(\beta_{nn} + \beta_{yn})$

Table 8

The case-only odds ratio in seven hypothetical populations

Population	A	B	C	D	E	F	G
β_{yy}/τ	0.01	0.01	0.01	10^{-6}	0.01	0.01	0.01
β_{yn}/τ	0.00	0.00	0.00	10^{-6}	10^{-6}	0.005	0.0045
β_{nn}/τ	0.01	0.01	0.001	0.01	10^{-6}	0.015	0.01
δ_y/τ	0.01	0.00	0.21	0	0	0	0.2
δ_n/τ	0.00	0.01	0.01	0	0	0	0.3
κ	2	$\frac{1}{2}$	2	2	2	2	1
$(\beta_+ + \delta_+)/\beta_+$	1.5	1.5	21	1	1	1	21.4

The odds ratio in Table 7 does not involve θ ; it is the case-only odds ratio κ given by

$$\begin{aligned} \kappa &= \frac{(\beta_{yy} + \beta_{yn} + \delta_y)(\beta_{nn} + \beta_{yn})}{(\beta_{nn} + \delta_n)\beta_{yy}} \\ &= \left(\frac{\beta_{yy} + \beta_{yn} + \delta_y}{\beta_{yy}} \right) \left(\frac{\beta_{nn} + \delta_n}{\beta_{nn} + \beta_{yn}} \right)^{-1}. \end{aligned} \tag{2}$$

Although κ can be consistently estimated in a case-only study derived from a randomized experiment, its magnitude lacks a clear causal interpretation. Certainly, expression (2) does not suggest one; unlike (1) for inherited genes and (2) is not a ratio of relative risks.

To begin, consider the hypothetical populations in Table 8, which presents seven populations, their case-only odds ratios, κ , and the overall relative risk of cancer from exposure to the agent, $(\beta_+ + \delta_+)/\beta_+$. In populations A and B, the exposure to the agent increases the risk of cancer by $1.5 = (\beta_+ + \delta_+)/\beta_+$ or 50%, but $\kappa = 2$ in A and $\kappa = 1/2$ in B, because in A the cancers caused by the agent have the mutation whereas in B they do not. In B, the agent never causes the mutation. If population C were exposed to the agent, the rate of cancer would be 21 times greater than without exposure, and 11 out of every 12 cancers would contain the mutation under study, but $\kappa = 2$. In populations D, E, and F, no cancers are caused by exposure to the agent, yet $\kappa = 2$. In D, about 1% of the population develops tumors without mutations whether exposed to the agent or not, about two people per million have tumors with the mutation under study, although none of the cancers are caused by the agent. In E, most tumors contain the mutation, but none of the tumors were caused by the agent. In population F, the agent does not cause cancer, but causes some mutations, and $\kappa = 2$. In contrast, in population G, the agent is very active in causing cancers and mutations, yet $\kappa = 1$. In short, if everyone in the population were exposed, the agent would cause 1/3 of the cancers in A, 11/12 of the cancers in C, and none of the cancers in D, E, and F; yet in all five populations, $\kappa = 2$. An investigator who wished to distinguish populations A, C, D, E, and F would have difficulty doing so using the case-only odds ratio, κ .

As a causal parameter, κ is ambiguous, not irrelevant. It does provide information. In Table 5, in terms of the cancer and mutation under study, $\nu + \beta_{nn} + \beta_{yy}$ people are entirely unaffected by exposure to the agent, and $\beta_{yn} + \delta_y + \delta_n$ are

affected in some way. Under the independence assumption, if $\kappa \neq 1$, then exposure causes something, that is, logically

$$(\kappa > 1) \implies (\beta_{yn} > 1) \vee (\delta_y > 0),$$

$$(\kappa < 1) \implies (\delta_n > 0),$$

but, alas, these implications are not equivalences. For instance, it is possible that $\delta_y > 0$, so exposure causes some tumors with mutations, yet $\kappa = 1$ or $\kappa < 1$. Ambiguity may diminish with information from other sources. For instance, there is no real doubt that K-ras mutations play an important role in the development of some cancers, so *if the independence assumption were true* in Section 1.2, then rejecting $H_0: \kappa = 1$ in favor of $\kappa > 1$ would provide some evidence that coffee causes some K-ras mutations in pancreatic tumors.

3.5 A Ratio of Risk Ratios

Some additional insight into the behavior of κ is possible by contrasting it with a certain, fairly unusual, ratio of risk ratios. If it happened to be true that $\beta_{yn} = 0$, then the case-only odds ratio κ would equal a ratio of risk ratios, namely:

$$\rho = \left(\frac{\beta_{yy} + \delta_y}{\beta_{yy}} \right) \left(\frac{\beta_{nn} + \delta_n}{\beta_{nn}} \right)^{-1}, \quad (3)$$

where $(\beta_{yy} + \delta_y)/\beta_{yy}$ would be the relative risk among individuals who would have a tumor with the mutation if exposed to the agent, and $(\beta_{nn} + \delta_n)/\beta_{nn}$ would be the relative risk among individuals who would have a tumor without the mutation if exposed. Now, κ in (2) can differ from 1 even if $\delta_y = \delta_n = 0$, that is, even if exposure never causes the cancer under study, but $\rho \neq 1$ implies either $\delta_y \neq 0$ or $\delta_n \neq 0$ or both. If $\beta_{yn} > 0$, then $\kappa > \rho$. Unfortunately, it is not safe to assume that $\beta_{yn} = 0$. There are two issues: multiple mutational pathways to the same clinical type of tumor, and genomic instability.

A person is counted in β_{yn} if this person would have developed a tumor of the clinical type under study whether exposed to the agent or not, but the tumor would have contained the mutation under study only if the individual were exposed. That is, the agent caused this person to have a tumor with the mutation rather than a tumor without the mutation, but the person would have had a tumor of this clinical type in either case. For instance, Section 3.2 speculated about the effects of work with asbestos on a hypothetical smoker: Perhaps he would develop a lung tumor whether exposed to asbestos or not, but perhaps the addition of asbestos exposure would cause an earlier, different tumor of the same clinical type, beginning in a different cell. Here, the two different tumors of the same clinical type may be produced by different sequences of mutations.

To be specific, consider the complementary roles proposed by Rajagopalan et al. (2002) for K-ras and B-raf mutations in colorectal cancers. They note that raf genes encode kinases regulated by ras genes; that is, different genes involved in the same process of growth regulation. They observed that in 330 colorectal tumors, 32 had B-raf mutations and 169 had K-ras mutations, but none had both. This would be expected if either a B-raf mutation or a K-ras mutation suffices to disrupt this aspect of growth regulation, and if both mutations are rare. If B-raf mutations and K-ras mutations were caused by different environmental agents, then preventing an exposure

that caused a B-raf mutation in one cell might prevent that cell from developing into a colorectal tumor, but it would not be expected to prevent K-ras mutations in other cells, so the same person might develop a colorectal tumor with a K-ras mutation anyway. If the mutation under study is not essential for the clinical type of tumor, and exposure sometimes causes the mutation, one expects $\beta_{yn} > 0$.

Genomic instability is reviewed by Lengauer et al. (1998), Orr-Weaver and Weinberg (1998), and Loeb (2001), and it has been incorporated into probability models for carcinogenesis by, for example, Little and Wright (2003). Genomic instability is the hypothesis that premalignant cells are more prone to mutation, or less able to resist or repair genomic damage, than are normal cells. The hypothesis is used to explain the diversity and large number of mutant genes found in tumors, and various cellular mechanisms have been suggested, including failures of DNA repair and failures during cell division resulting in aneuploidy. It is possible that a certain mutagen rarely if ever damages normal cells, but is highly mutagenic in a cell in which certain DNA repair mechanisms have been disabled. If a mutagen is more likely to damage the cells in a colony of premalignant cells than to damage a normal cell, β_{yn} might not be small compared to δ_y .

In short, it is not safe to assume a priori that $\beta_{yn} = 0$ and $\kappa = \rho$.

4. Summary and Discussion

Under the independence assumption, the case-only odds ratio (1) measures interaction, on the risk ratio scale, between an environmental agent and inherited genes in the causation of disease. In contrast, with mutations, the magnitude of this odds ratio, κ in (2), does not have a clear, causal interpretation.

The difference between inherited genes in Table 3 and mutations in Table 5 is that the environmental agent may cause mutations but does not cause the inheritance of particular genes. In the language of experimental design, inherited genes are covariates, that is, determined prior to exposure to the agent and therefore unaffected by agent; whereas mutations are an outcome, that is, determined after exposure to the agent and possibly affected by exposure. The same logic applies to any observed binary covariate or outcome independent of exposure—it need not be genetic in nature. That is, for risk ratios, the case-only design may be used to study interaction between a covariate and an agent when the independence assumption, defined in Section 2, holds with that covariate; the covariate need not be an inherited gene, but it must not be affected by the exposure.

With inherited genes, most discussions of the independence assumption are formally explicit about the independence of inherited genes and environmental exposures, but are informal about the absence of confounding due to unmeasured covariates. In contrast, as discussed here, the independence assumption formally combines the usual independence assumption with strong ignorability, in the sense of Rosenbaum and Rubin (1983), which is a formal expression of the absence of hidden bias due to unobserved covariates. As with other observational studies, in a case-only study that compares exposed cases to unexposed cases, confounding of exposure with unobserved covariates is certainly possible. For instance, coffee drinkers and abstainers may have different diets or habits.

Table 9

Checking the independence assumption in a case-control study: expected counts for the control-only odds ratio under the independence assumption

	Exposed	Not Exposed
Genotype A	$\theta\nu_A$	$(1 - \theta)(\nu_A + \delta_A)$
Genotype B	$\theta\nu_B$	$(1 - \theta)(\nu_B + \delta_B)$

In a case-control study that includes a case-only analysis with inherited genes, Botto and Khoury (2001) suggest also examining the “control-only” odds ratio as a check on the independence assumption. Using Table 3, the control-only odds ratio under the independence assumption is the odds ratio from the expected counts in Table 9, which is the complement of Table 4. If the disease is extremely rare for both genotypes, so that $\nu_A \gg \delta_A$ and $\nu_B \gg \delta_B$, then the control-only odds ratio in Table 9 is approximately 1, which is Botto and Khoury’s suggested check on the independence assumption. Although Stern et al. (2002) reported case-only analyses, they also had controls who were urology patients without a history of cancer “frequency matched to cases based on ethnicity, sex, and age at interview.” In their data, the sample odds ratio of 1.76 does not differ significantly from 1 by Fisher’s exact test, with one-sided significance level 0.13, so the independence assumption is not sharply contradicted by the data from controls.

ACKNOWLEDGEMENTS

This work has been supported by a grant from the U.S. National Science Foundation and grants R01-CA95664 and R01-CA98901 from the U.S. National Cancer Institute.

RÉSUMÉ

Dans les études cas-cas les plus simples, les malades sont classés dans une table 2*2 en fonction d’une caractéristique génotypique et de l’exposition à un agent environnemental. Dans certaines instances, la caractéristique génotypique est fonction d’un allèle hérité; Dans d’autres elle est fonction de mutations non héritées affectant par exemple des proto-oncogènes ou des gènes suppresseur de tumeur favorisant le développement d’un cancer. Ici, l’odds ratio cas-cas s’écrit comme un paramètre causal en termes de conséquence potentielle selon la présence ou l’absence d’exposition à l’agent environnemental. Nous montrons que cet odds ratio peut s’interpréter comme un paramètre causal dans le cadre d’allèles hérités, mais que sa valeur n’a pas d’interprétation causale lorsqu’il s’agit de mutations non héritées, même si une déviation par rapport à 1 fournit une information. La différence vient de ce que l’agent environnemental n’a certainement pas causé le fait qu’un individu hérite d’un allèle particulier tandis qu’il a pu causer la mutation non héritée.

REFERENCES

Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology* **154**, 687–693.

Amoguera, C., Shibata, D., Forrester, K., Martin, J., Arnhem, N., and Perucho, M. (1988). Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes. *Cell* **53**, 549–554.

Begg, C. and Zhang, Z. (1994). Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiology, Biomarkers and Prevention* **3**, 173–175.

Botto, L. D. and Khoury, M. J. (2001). Commentary: Facing the challenge of gene-environment interaction: The two-by-four table and beyond. *American Journal of Epidemiology* **153**, 1016–1020.

Cochran, W. G. (1965). The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society, Series A* **128**, 134–155.

Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health* **43**, 677–683.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Hamilton, M. A. (1979). Choosing the parameter for 2×2 or $2 \times 2 \times 2$ table analysis. *American Journal of Epidemiology* **109**, 362–375.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* **100**, 57–70.

Khoury, M. J. and Flanders, W. D. (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: Case-control studies with no controls! *American Journal of Epidemiology* **144**, 207–213.

Lehmann, A. R. (2001). The xeroderma pigmentosum group D (XPD) gene: One gene, two functions, three diseases. *Genes and Development* **15**, 15–23.

Lengauer, C., Kinzler, K. W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature* **396**, 643–649.

Little, M. P. and Wright, E. G. (2003). A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Mathematical Biosciences* **183**, 111–134.

Loeb, L. A. (2001). A mutator phenotype in cancer. *Cancer Research* **61**, 3230–3239.

McKinnel, R. G., Parchment, R. E., Perantoni, A. O., and Pierce, G. B. (1998). *Biological Basis of Cancer*. New York: Cambridge.

NCI. (1999). Innovative study designs and analytic approaches to the genetic epidemiology of cancer. *Journal of the National Cancer Institute Monographs* **26**, 1–105.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9 (In Polish), *Roczniki Nauk Roimicznych* **X**, 1–51, Reprinted in English with Discussion in *Statistical Science*, 1990, **5**, 463–480.

Neyman, J. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society, Series B* **2**, 107–180.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194**, 23–28.

Orr-Weaver, T. L. and Weinberg, R. A. (1998). A checkpoint on the road to cancer. *Nature* **392**, 223–224.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13**, 153–162.

Porta, M., Malats, N., Guarner, L., et al. (1999). Association between coffee drinking and K-ras mutations in exocrine

- pancreatic cancer. *Journal of Epidemiology and Community Health* **53**, 702–709.
- Potter, J. D. (2002). Pancreas cancer—we know about smoking, but do we know anything else? *American Journal of Epidemiology* **155**, 793–795.
- Rajagopalan, H., Bardelli, A., Lengauer, C., Kinzler, K. W., Vogelstein, B., and Velculescu, V. E. (2002). RAF/RAS oncogenes and mismatch repair status. *Nature* **418**, 934.
- Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88**, 219–231.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.
- Schmidt, S. and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene–environment interaction. *American Journal of Epidemiology* **150**, 878–885.
- Smith, J. M. (1968). *Mathematical Ideas in Biology*. New York: Cambridge University Press.
- Stern, M. C., Johnson, L. R., Bell, D. A., and Taylor, J. A. (2002). XPD Codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. *Cancer Epidemiology, Biomarkers and Prevention* **11**, 1004–1011.
- Stolzenberg-Solomon, R. Z., Pietinen, P., Taylor, P. R., Virtamo, J., and Albanes, D. (2002). Prospective study of diet and pancreatic cancer in male smokers. *American Journal of Epidemiology* **155**, 783–792.
- Weinberg, C. R. and Umbach, D. M. (2000). Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *American Journal of Epidemiology* **152**, 197–203.

Received December 2002. Revised July 2003.

Accepted September 2003.

APPENDIX

Carcinogenesis Models and Exposure Effects

A reader interested in carcinogenesis models might wish to relate them to Table 5. In principle, any such model will gen-

erate a probability distribution in Table 5. To provide a very brief, tangible, and simple illustration, consider essentially a one-hit model (Smith, 1968, Chapter 6), in which every individual is observed for the same t years and has the same number of cells. Cells are fired upon at a constant rate, and as soon as it is hit the cell becomes a recognizable cancer. If not exposed to the agent, the time to cancer from background alone is T_B which has an exponential distribution with hazard rate λ_B ; however, without exposure, the mutation under study never occurs, and the cancer is caused through some other form of damage to the genome. If exposed to the agent, the time to cancer is $T = \min(T_B, T_E)$, where T_B is as before, and T_E has an exponential distribution with hazard λ_E and is independent of T_B ; moreover, if and only if $T = T_E$ is the mutation present in the tumor. Write $\omega = \lambda_E/(\lambda_B + \lambda_E)$, $\pi_E = \exp(-t\lambda_E)$, and $\pi_B = \exp(-t\lambda_B)$. Then a person would not develop cancer over t years whether exposed or not if $T > t$ which happens with probability $\nu = \pi_E \times \pi_B$. A person would develop cancer only if exposed if $T_E \leq t < T_B$, and in this case the person has a cancer with the mutation, and this happens with probability $\delta_y = \pi_B(1 - \pi_E)$. A person would develop the cancer without the mutation whether exposed or not exposed if $T_B < T_E \leq t$ or $T_B \leq t < T_E$, which happens with probability $\beta_{nn} = (1 - \pi_B)(1 - \pi_E)(1 - \omega) + \pi_E(1 - \pi_B)$. A person would develop the cancer with the mutation if exposed and would develop the cancer without the mutation if not exposed if $T_E < T_B \leq t$, which happens with probability $\beta_{yn} = (1 - \pi_B)(1 - \pi_E)\omega$. By assumption $\delta_n = 0$, because if the exposure causes a cancer that would not occur without exposure, then $T_E \leq t < T_B$, but in this case the mutation is present. Also by assumption $\beta_{yy} = 0$, because mutations only occur in cancers caused by the agent. In this naive model, the fact that $\beta_{yy} = 0$ makes κ infinite, and independence makes β_{yn} strictly positive but small. In particular, κ provides no information about the parameters of the model, λ_B and λ_E . For instance, with $t = 1$, $\lambda_B = 0.7$, $\lambda_E = 10^{-6}$, about half the population would develop the cancer from background alone, whereas about one in a million would develop cancer from exposure without background, yet $\kappa = \infty$. The one-hit model permits brief illustration. Realistic models postulate repeated hits upon a growing colony of increasingly damaged cells (Little and Wright, 2003) yielding a different distribution for Table 5.