American Journal of
EPIDEMIOLOGY

Volume 137
Number 1
January 1, 1993

Sponsored by the Society for Epidemiologic Research

REVIEWS AND COMMENTARY

Toward a Clearer Definition of Confounding

Clarice R. Weinberg[1]

Epidemiologists are aware that the estimated effect of an exposure can be biased if the investigator fails to adjust for confounding factors when analyzing either a prospective or retrospective etiologic study. Standard texts warn, however, that intervening factors are an exception: one should not adjust for any factor which is intermediate on the causal pathway between the exposure and the disease. Other factors which are not on the causal pathway but are caused in part by the exposure are often adjusted for in epidemiologic studies. This paper illustrates that bias can result when adjustment is made for any factor which is caused in part by the exposure under study and is also correlated with the outcome under study. Intervening variables are only one example of this phenomenon. The misleading effects of this practice are illustrated with examples. *Am J Epidemiol* 1993;137:1–8.

bias (epidemiology); case-control studies; cohort studies; confounding factors (epidemiology); effect modifiers (epidemiology); epidemiologic methods; research design

It is well known that a measure of etiologic association between an exposure and risk of disease can be biased in an epidemiologic study whenever a third factor is associated with the exposure and also related to the risk of disease in the unexposed. Such a "confounding" factor can result in severe distortions if the investigator fails to adjust for it carefully in the analysis or match on it in the design of the study.

In a purely descriptive study, whose intent is to estimate absolute or relative risk, as related to all potentially prognostic factors, confounding per se is not an important problem: one simply includes all measured risk factors in the predictive model. By contrast, we shall consider a study whose objective is to assess a possible etiologic association.

It is well recognized (e.g., 1–3) that, in the context of an etiologic study, any factor which is intermediate on the causal pathway between the exposure and the disease (figure 1a) should not be treated as a confounder and should not be accounted for in the

Received for publication January 24, 1992, and in final form August 10, 1992

[1] National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Reprint requests to Dr. Clarice R. Weinberg, MD B3-02, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709

The author thanks the following persons for their comments on earlier versions of the manuscript: Drs. D Baird, N. Breslow, B. Gladen, S. Greenland, R. Little, N. Mantel, B. McKnight, K. Meier, J Ratcliffe, K. Rothman, S Wacholder, N Weiss, E. White, and A J Wilcox

analysis or matched on in the design. The purpose of this paper is to discuss a more general situation, in which a factor that changes the estimated effect of the exposure, and thus appears to be a confounder or an effect modifier, should in many instances be omitted from the analysis: namely, whenever the factor may itself have been affected by the exposure. Inclusion of such a factor, either as a matching factor, an explicit term in a model, or as a basis for stratification, can severely bias measures of exposure-associated risk, as will be illustrated.

Epidemiologists generally recognize that factors which are related to the disease outcome *only because* they are causally related to the exposure (e.g., membership in Alcoholics Anonymous among alcohol consumers) should not be considered confounders and should not be adjusted for (1). This paper goes beyond this rule to assert that bias can result even when a factor which is caused in part by the exposure is related to the disease among the unexposed. This means that, under certain circumstances, independent risk factors which appear to be confounders should not be adjusted for.

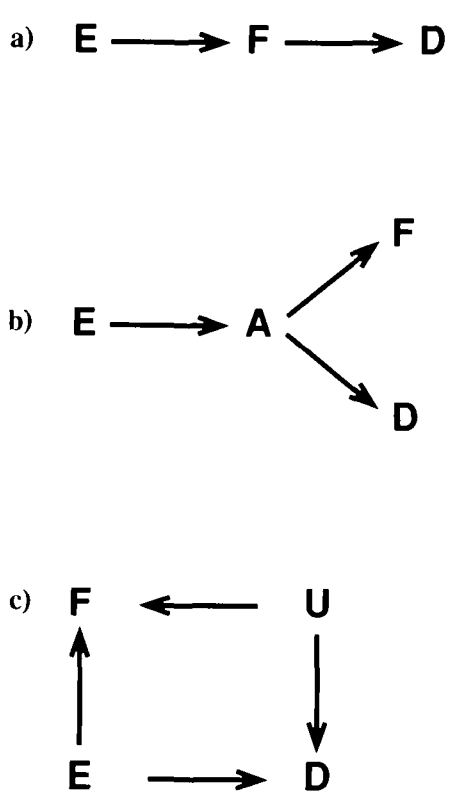

FIGURE 1. Three mechanisms by which factor *F* will be caused in part by exposure *E* and also associated with disease *D* in the unexposed: a) *F* is an intervening factor in the causal pathway between *E* and *D*; b) *F* and *D* are caused in part by a single underlying abnormality, *A*, which is caused in part by *E*; and c) *D* and *F* are both caused in part by *E*, and there is an unmeasured factor, *U*, which is perhaps genetic and which influences susceptibility to both.

EXAMPLE: SPONTANEOUS ABORTION HISTORY

Suppose a cohort of women is recruited early in pregnancy and followed for the occurrence of spontaneous abortion, where some but not all of these women report an exposure thought possibly to cause pregnancy loss. History of spontaneous abortion is known to be related to risk, even among the unexposed, and, under the study hypothesis, could also be related to the exposure, if the exposure tends to have been long term. Thus, it could be regarded as a potential confounder. At the same time, history of spontaneous abortion is almost certainly not on the causal pathway between exposure and loss of the current pregnancy. Such a history is more plausibly viewed as a marker for elevated risk (4), and this elevated risk may be due in part to effects of the same exposure on past pregnancies. History of spontaneous abortion does satisfy classical criteria for "confounding" given in standard epidemiology texts (e.g., see Rothman (3)): it presumably serves as a marker for causally related factors, and it may also be associated (under the study hypothesis) with the exposure under study. Should it then be adjusted for in the estimation of the exposure-associated risk?

For notational simplicity, let us suppose we have information for one additional prior pregnancy for each woman, and that the exposure of interest was either present during both pregnancies or absent during both. Suppose the exposure, E, acts by increasing the risk of a certain underlying abnormality, A, and that this abnormality was either present during both pregnancies or absent during both. (The abnormality might be, for ex-

ample, poor response of the endometrium to endogenous estrogens.) Let S_1 denote the loss of the first pregnancy and S_2 the loss of the second. Assume the two outcomes are independent, conditional on the presence or absence of the (unobservable) abnormality. If we also assume, for simplicity, that the risk conditional on both A and E status in fact depends only on A, then we can write:

$$\begin{aligned}\Pr[S_1|E] &= \Pr[S_1|A, E]\Pr[A|E] \\ &\quad + \Pr[S_1|\bar{A}, E]\Pr[\bar{A}|E] \\ &= \Pr[S_1|A]\Pr[A|E] \\ &\quad + \Pr[S_1|\bar{A}]\Pr[\bar{A}|E].\end{aligned}$$

For instance, suppose the risk of the abnormality, A, is 0.385 for the exposed and 0.15 for the unexposed, and the risk of spontaneous abortion is 0.9 for those with A and 0.1 for those without A. Applying the above expression, the risk would be $(0.9)(0.385) + (0.1)(0.615) = 0.408$ per exposed pregnancy, and 0.22 per unexposed pregnancy, for a true relative risk of 1.85.

Now suppose we stratify on history of spontaneous abortion. For exposed women whose earlier pregnancy spontaneously aborted, the relative risk is developed as follows:

$$\begin{aligned}\Pr[S_2|E, S_1] &= \frac{\Pr[S_1, S_2|E]}{\Pr[S_1|E]} \\ &= \frac{\Pr[S_1, S_2|A]\Pr[A|E] + \Pr[S_1, S_2|\bar{A}]\Pr[\bar{A}|E]}{\Pr[S_1|E]} \\ &= \frac{(0.9)^2(0.385) + (0.1)^2(0.615)}{0.408} \\ &= 0.779.\end{aligned}$$

A similar calculation for the unexposed yields a conditional risk of 0.591, for a stratum-specific relative risk of 1.32. The calculations for the stratum with a prior live birth also lead to a stratum-specific relative risk of 1.32. Thus, there is homogeneity across the strata defined by history of spontaneous abortion, and apparent confounding, in a mathematical sense, since the pooled relative risk following stratification is markedly lower than the crude relative risk, which was 1.85.

The analysis that "adjusts" for history of spontaneous abortion has thus markedly reduced the estimated exposure-related relative risk. And yet, under our assumptions, if the exposed women had never been exposed, their risk would not differ from that of the unexposed women, so that, by recently proposed definitions (5), there is no confounding to be adjusted for. It seems clear, then, that the inclusion of this factor as a stratification variable has adjusted away part of the effect under study and understated the importance of the exposure in increasing the risk of spontaneous abortion.

Breslow and Day (1) warned of a similar bias toward the null in a related context where there has been "overmatching." In their examples, stratification was carried out on a medical condition which revealed signs of early stages of the disease under study, such as chronic cough when lung cancer is the endpoint, or uterine bleeding when endometrial cancer is the endpoint of interest.

A modification of the spontaneous abortion example shows that, in our more general context, the bias need not be toward the null, and the stratum-specific relative risks need not be equal. Consider instead what happens if the risk of the abnormality is 0.75 among those with the exposure, and 0.3 among those without. If the risk of spontaneous abortion is 0.7 in those with the abnormality and 0.05 in those without, then the true exposure-associated relative risk can be shown to equal 2.19. Among women whose prior pregnancy ended in spontaneous abortion, the relative risk is 1.13, while, among women without such a history, the relative risk is 2.87. The investigator could report these estimates separately, interpreting the discrepancy as strong evidence for effect modification. For example, it might be argued that those without a history of spontaneous abortion are less likely to have a genetically based problem and more likely to have an environmental cause, and so on. Again, the inclusion of the correlated factor has effectively muddled the assessment of exposure-associated risk.

A skeptic might argue that the case here is not so clear and that a good reason for including a factor such as history of spontaneous abortion is that it can serve to some extent as a surrogate for unmeasured risk factors that may themselves be confounders. While it is true that those potential confounders being implicitly adjusted for inappropriately include the exposure under study, couldn't they also include some genuine confounders?

One sees readily, however, that such an adjustment does a remarkably poor job of "controlling" for unmeasured confounders. Consider a scenario identical to the one just considered, except that the exposure is unrelated to risk, while some bad actor, C, invariably accompanies it. Then, the above algebraic development yields a relative risk of 1.13 among those with a prior loss and 2.87 among those without, results quite inconsistent with the truth (1.0). Thus, considered naively as an adjustment for unmeasured confounding variables, that strategy is appallingly ineffective.

GENERAL PROBLEM

More generally, suppose disease D is under study and data are available on the presence or absence of some factor, F. Suppose the exposure, E, increases the risk of F and of D. Suppose also that F and D are correlated among those with the exposure. Such a correlation could arise, for example, when F and D are separate manifestations of the same underlying abnormality (figure 1b), as in the preceding example. Examples are easy to find. In a recent study to assess the role of blood pressure in cognitive impairment in the elderly, use of cardiovascular medications, self-reported history of stroke, and self-assessment of health were all listed as potential confounders (6). The use of medications is influenced by the "exposure" under study (blood pressure), history of stroke may be intermediate on the causal pathway, and self-assessment of poor health may be influenced both by the exposure and, what may be worse, by the endpoint under study. Results corresponding to those illustrated above will again hold if such factors are treated as confounders, leading to misleading conclusions regarding the role of blood pressure in cognitive impairment.

What happens in the situation where the exposure increases the risk of both F and D, but the two are independent within the exposed and within the unexposed? One can show by straightforward algebra that if D and F occur independently within the exposed and also within the unexposed, then the stratum-specific relative risks correspond to the crude, pooled-data relative risk. (This also follows from the classical understanding of confounding: a factor must be associated with the disease among the unexposed to be a confounder.) Thus the problem arises only when F is associated with D within at least one of the strata defined by the exposure.

EXAMPLES OF CORRELATION-INDUCING SCENARIOS

It is instructive to consider three specific scenarios where this kind of exposure-related factor would arise. First, F and D may represent separate manifestations of a single underlying abnormality (figure 1b), which is caused by E, as in the spontaneous abortion example. Second, there could be correlated susceptibility to F and D among members of the population, due to unmeasured shared risk factors (figure 1c), e.g., genetic. Or, third, a correlation between F and D may arise secondary to variations in true exposure within exposure strata, provided risks of F and D are both dose-related to the exposure. In a realistic situation, any mix of these three mechanisms is possible.

First, suppose, as in the spontaneous abortion example, that F and D are separate manifestations of the same abnormality and the risk of that abnormality is affected by the exposure, E (figure 1b). Suppose that $\Pr[F, D|E, A] = \Pr[F, D|A]$, where A denotes the underlying abnormality. Thus, we assume for simplicity that E affects the outcomes F and D only through A. One can show (Appendix 1) that extreme bias toward the null can result from stratification on the correlated factor, F. The spontaneous abor-

tion example also illustrated that stratification on F can induce heterogeneity among risk ratios.

Second, suppose the risk of F, $\Pr_i(F)$, and the risk of D, $\Pr_i(D)$, covary among individuals, indexed by i, in the population (e.g., due to common causal components which have not been measured), but the outcomes are independent within each individual. As an example of this, rehabilitated intravenous drug users may be prone to other risky behaviors as well and may be at increased risk for accidental death in later life. As another example, women with human papillomavirus may also be susceptible to other venereal pathogens, due to inherent variations in resistance among women with similar exposures.

We suppose that, for each individual, the joint occurrence probability is given as the product, $\Pr_i(D)\Pr_i(F)$ (i.e., the outcomes are independent within individuals), but of course the individual risks, $\Pr_i(D)$ and $\Pr_i(F)$, are unobservable. For simplicity of illustration, assume a study where a cohort is followed for a fixed length of time, and one ascertains whether or not each member has developed F and D by the end of follow-up.

The bias induced here by stratification on F can be dramatic. If one assumes that the exposure-associated relative risk for D is R_D for each individual, then the estimated relative risk is R_D for the F stratum, but is not R_D for the stratum without F (see Appendix 2). For that stratum, the relative risk can be biased either toward or away from the null, and this bias can be extreme. Thus, under this scenario, considerable bias, taking the form of apparent effect modification, can arise from stratifying on the correlated factor, F.

Third, suppose the measured exposure E has to some extent been misclassified, or a number of different exposure levels have been grouped, and F and D both have a positive dose-response relation to the true exposure, but occur independently at each level of true exposure. One can show (Appendix 3) that the apparent relative risk within the stratum including those with F is elevated compared to the truth. The best way to intuit this is that among those with F the true exposure of the exposed is relatively high. Correlated, exposure-related factors may be common. Even if D and F are physiologically distinct endpoints, if there is a dose-related effect of the exposure on both, then correlation between them will be induced even with the best dosimetry, if there are person-to-person differences in absorption of E or in the efficacy of relevant metabolic detoxification pathways.

DISCUSSION

While it is recognized that, when the exposed (such as those undergoing estrogen therapy) come under increased medical surveillance, one cannot hope to correct the resulting disease detection bias by stratifying on the diagnostic procedures to which study participants have been subjected (7), the general problem of apparent confounding or effect modification due to consequences of exposure seems to have been passed over in the standard epidemiology texts (1–3, 8). The present development has shown that bias can result from adjusting for or matching on such a factor, even when the factor is not intermediate on the causal pathway.

While the development here has been limited to a dichotomous other factor and a dichotomous exposure, the problem occurs in broader settings as well. The "factor" could be the results of a laboratory test, such as serum cholesterol level, and the exposure could be on a continuous scale; the same concerns would apply. One way to look at this problem is that the correlated factor, F, serves as a partial surrogate for the disease, so that the endpoint is to some extent represented on both sides of the model equation.

Examples of factors causally related to the exposure under study are plentiful in the epidemiologic literature. Some fall straightforwardly into the paradigm we have considered (6, 9, 10). Other examples are more subtle. One study matched drivers involved in two-car collisions, defining the at-fault driver as the case, to assess the role of alcohol in fatal auto accidents (11). In this study,

the factor matched on was involvement in a fatal accident, and this could also be causally related to the exposure (alcohol), regardless of which driver was judged to be at fault. (People driving under the influence are less able to avoid accidents and also less able to respond appropriately once an accident has occurred.)

The basis for the association between a potential confounding factor and the disease under study may often be unclear, and informed decisions must be made. If history of infertility is considered when ovarian cancer is the outcome of interest, such a history could plausibly play a double role, both serving as a marker for inherent risk, and simultaneously as a manifestation of an underlying abnormality that may have been caused in part by the exposure under study, if that exposure tended to have been long term. In my view, the model that does not include adjustment for the correlated endpoint is the more credible one when the basis for the exposure-"confounder" relation is ambiguous.

Epidemiologists are acutely conscious of the danger of over-interpreting associations as causal, and it may be as a consequence of this that they sometimes avoid thinking about the potentially causal nature of associations between exposures of interest and potential confounders. It is all too easy to fall into a purely empirical approach to analysis, where covariates are added to the model one by one and retained if they seem to make a difference. Valid inference would be better served if, perhaps with the aid of causal diagrams, careful consideration were given to whether each factor should be in the model, particularly if the factor may have been caused in part by the exposure under study. The standard definitions of confounding should be revised to specify that factors which may have been caused in part by the exposure under study should not routinely be treated as potential confounders.

REFERENCES

1. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. (IARC Scientific Publications No. 32). Lyon: International Agency for Research on Cancer, 1980.
2. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.
3. Rothman KJ. Modern epidemiology. Boston: Little, Brown & Co, 1986.
4. Gladen BC. On the role of "habitual aborters" in the analysis of spontaneous abortion. Stat Med 1986;5:557–64.
5. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986;15:413–19.
6. Scherr PA, Hebert LE, Smith LA, et al. Relation of blood pressure to cognitive function in the elderly. Am J Epidemiol 1991;134:1303–15.
7. Greenland S, Neutra R. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial concer. J Chronic Dis 1981; 34:433–8.
8. Schlesselman JJ. Case-control studies: design, conduct, analysis. New York: Oxford University Press, 1982.
9. Grove JS, Nomura A, Severson RK, et al. The association of blood pressure with cancer incidence in a prospective study. Am J Epidemiol 1991;134: 942–7.
10. Li D-K, Daling JR. Maternal smoking, low birth weight, and ethnicity in relation to sudden infant death syndrome. Am J Epidemiol 1991;134: 958–64.
11. Perneger T, Smith GS. The driver's role in fatal two-car crashes: a paired "case-control" study. Am J Epidemiol 1991; 134:1138–45.

APPENDICES

APPENDIX 1

The case where D and F are both manifestations of the same underlying abnormality, A. We can write the following decomposition:

$$\Pr[F, D|E] = \Pr[F, D|E, A]\Pr[A|E] + \Pr[F, D|E, \bar{A}]\Pr[\bar{A}|E].$$

We can rewrite:

$$\frac{\Pr[D|E, F]}{\Pr[D|\bar{E}, F]} = \frac{\Pr[F, D|E]\Pr[F|\bar{E}]}{\Pr[F, D|\bar{E}]\Pr[F|E]}.$$

We can, with substitution, expand this as follows:

$$\frac{\{[\Pr[F, D|A]\Pr[A|E] + \Pr[F, D|\bar{A}]\Pr[\bar{A}|E]\}\{\Pr[F|A]\Pr[A|\bar{E}] + \Pr[F|\bar{A}]\Pr[\bar{A}|\bar{E}]\}}{\{\Pr[F, D|A]\Pr[A|\bar{E}] + \Pr[F, D|\bar{A}]\Pr[\bar{A}|\bar{E}]\}\{\Pr[F|A]\Pr[\mathrm{A}|\mathrm{E}] + \Pr[F|\bar{A}]\Pr[\bar{A}|E]\}}$$

$$= \frac{\{[\Pr[F, D|A]R\epsilon + \Pr[F, D|\bar{A}](1 - R\epsilon)\}\{\Pr[F|A]\epsilon + \Pr[F|\bar{A}](1 - \epsilon)\}}{\{\Pr[F, D|A]\epsilon + \Pr[F, D|\bar{A}](1 - \epsilon)\}\{\Pr[F|A]R\epsilon + \Pr[F|\bar{A}](1 - R\epsilon)\}},$$

where ϵ denotes the risk of A in the unexposed and R is the exposure-associated relative risk for A. This relative risk (and the analogous one for those without F) can be very close to 1.0 when ϵ is small. (As we pass to the limit, letting ϵ approach 0, the ratio tends to 1.) Thus, extreme bias toward the null can result from stratification on the correlated factor, F.

APPENDIX 2

The case where susceptibilities to D and F are correlated among individuals in the population. We suppose that for each individual the joint occurrence probability is given as the product $\Pr_i(D)\Pr_i(F)$ (i.e., the outcomes are independent within individuals), but of course the individual risks $\Pr_i(D)$ and $\Pr_i(F)$ are unobservable. For simplicity of illustration, assume a study where a cohort is followed for a fixed length of time, and one ascertains whether or not each member has developed F and D by the end of follow-up. Under this design, the joint occurrence probability for a randomly selected member of the population is $\Pr(F, D) = \mathrm{Ex}[\Pr_i(F)\Pr_i(D)]$, where the notation Ex[ ] denotes averaging across members of the population. Suppose the exposure occurs independently of inherent risk, and affects the risk of D for each individual by a factor R_D, and the risk of F by a factor R_F. Then the joint probability for the two outcomes in a randomly selected exposed individual is given by:

$$\Pr(F, D|E) = R_D R_F \mathrm{Ex}[\Pr_i(F|\bar{E})\Pr_i(D|\bar{E})]$$

and the relative risk for the stratum with F is developed as follows:

$$\frac{\Pr D|E, F)}{\Pr(D|\bar{E}, F)} = \frac{\Pr(F, D|E)\Pr(F|\bar{E})}{\Pr(F, D|\bar{E})\Pr(F|E)} = \frac{R_D R_F \mathrm{Ex}[\Pr_i(F|\bar{E})\Pr_i(D|\bar{E})]\mathrm{Ex}[\Pr_i(F|\bar{E})]}{\mathrm{Ex}[\Pr_i(F|\bar{E})\Pr_i(D|\bar{E})]R_F\mathrm{Ex}[\Pr_i(F|\bar{E})]} = R_D.$$

Somewhat different results hold for the stratum without F:

$$\frac{\Pr(D|E, \bar{F})}{\Pr(D|\bar{E}, \bar{F})} = \frac{\Pr(\bar{F}, D|E)\Pr(\bar{F}|\bar{E})}{\Pr(\bar{F}, D|\bar{E})\Pr(\bar{F}|E)}$$

$$= R_D \frac{\{\mathrm{Ex}[\Pr_i(D|\bar{E})] - \mathrm{R}_F\mathrm{Ex}[\Pr_i(F|\bar{E})\Pr_i(D|\bar{E})]\}\{1 - \mathrm{Ex}[\Pr_i(F|\bar{E})]\}}{\{\mathrm{Ex}[\Pr_i(D|\bar{E})] - \mathrm{Ex}[\Pr_i(F|\bar{E})\Pr_i(D|\bar{E})]\}\{1 - \mathrm{R}_F\mathrm{Ex}[\Pr_i(F|\bar{E})]\}}.$$

If R_F is 1 there is no bias in this stratum; and if the risks for F and D do not covary among the members of the population, then there is no bias. But (positively) correlated susceptibility to F and D does produce a downward bias, when R_F is greater than 1.0, since the above multiplier of R_D can be shown to be less than 1 under those assumptions. If, on the other hand, the correlation between the risk for F and that for D is negative, then the bias is away from the null.This bias can be dramatic. For example, if the joint occurrence probability for the unexposed is half the risk for D alone and R_F is 2, then the above expression becomes 0.

APPENDIX 3

The case where correlation is secondary to imprecise exposure assessment. Again we begin with the general identity:

$$\frac{\Pr[D|E,\ F]}{\Pr[D|\bar{E},\ F]} = \frac{\Pr[F,\ D|E]\Pr[F|\bar{E}]}{\Pr[F,\ D|\bar{E}]\Pr[F|E]},$$

but we make use of the assumption that F and D are independent in the unexposed to rewrite this as:

$$\frac{\Pr[F,\ D|E]\Pr[F|\bar{E}]}{\Pr[F,\ D|\bar{E}]\Pr[F|E]} = \frac{\Pr[F,\ D|E]\Pr[F|\bar{E}]}{\Pr[F|\bar{E}]\Pr[D|\bar{E}]\Pr[F|E]} = \frac{\Pr[F,\ D|E]}{\Pr[F|E]\Pr[D|\bar{E}]}.$$

But

$$\frac{\Pr[F,\ D|E]}{\Pr[F|E]\Pr[D|\bar{E}]} > \frac{\Pr[D|E]}{\Pr[D|\bar{E}]},$$

since

$$\Pr[F,\ D|E] > \Pr[F|E]\Pr[D|E],$$

where the latter inequality follows from the positive correlation induced by the variation in true exposure within those with measured exposure E. Thus, the apparent relative risk within the stratum defined by the presence of F is elevated compared to the truth.