CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

1

Introduction

1.1 Why should I do multivariable analysis?

DEFINITION

Multivariable analysis is a tool for determining the relative contributions of different causes to a single event.

We live in a multivariable world. Most events, whether medical, political, social, or personal, have multiple causes. And these causes are related to one another. Multivariable analysis[1] is a statistical tool for determining the relative contributions of different causes to a single event or outcome.

Clinical researchers, in particular, need multivariable analysis because most diseases have multiple causes, and prognosis is usually determined by a large number of factors. Even for those infectious diseases that are known to be caused by a single pathogen, a number of factors affect whether an exposed individual becomes ill, including the characteristics of the pathogen (e.g., virulence of strain), the route of exposure (e.g., respiratory route), the intensity of exposure (e.g., size of inoculum), and the host response (e.g., immunologic defense).

Multivariable analysis allows us to sort out the multifaceted nature of risk factors and their relative contribution to outcome. For example, observational epidemiology has taught us that there are a number of risk factors associated with premature mortality, notably smoking, a sedentary lifestyle, obesity, elevated cholesterol, and hypertension. Note that I did not say that these factors *cause* premature mortality. Statistics alone cannot prove that a relationship between a risk factor and an outcome are causal.[2] Causality is established on

[1] The terms "multivariate analysis" and "multivariable analysis" are often used interchangeably. In the strict sense, multivariate analysis refers to simultaneously predicting multiple outcomes. Since this book deals with techniques that use multiple variables to predict a single outcome, I prefer the more general term multivariable analysis.

[2] Throughout the text I use the terms "associated with" and "related to" interchangeably. Similarly, I use the terms "risk factor," "exposure," "predictor," and "independent variable," and the terms "outcome" and "dependent variable," interchangeably. Although some of these terms such as "risk factor," "predictor," and "outcome" imply causality remember that causality can never be proven with statistical analysis. The best way for establishing causality is through rigorous study design (e.g., randomization to eliminate confounding, longitudinal observations to minimize the chance that the "outcome" caused the "risk factor").

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

the basis of biological plausibility and rigorous study designs, such as randomized controlled trials, which eliminate sources of potential bias.

Identification of risk factors of premature mortality through observational studies has been particularly important because you cannot randomize people to many of the conditions that cause premature mortality, such as smoking, sedentary lifestyle, or obesity. And yet these conditions tend to occur together; that is, people who smoke tend to exercise less and be more likely to be obese.

How does multivariable analysis separate the *independent* contribution of each of these factors? Let's consider the case of exercise. Numerous studies have shown that persons who exercise live longer than persons with sedentary lifestyles. But if the only reason that persons who exercise live longer is that they are less likely to smoke and more likely to eat low-fat meals leading to lower cholesterol, then initiating an exercise routine would not change a person's life expectancy.

The Aerobics Center Longitudinal Study tackled this important question.[3] They evaluated the relationship between exercise and mortality in 25,341 men and 7080 women. All participants had a baseline examination between 1970 and 1989. The examination included a physical examination, laboratory tests, and a treadmill evaluation to assess physical fitness. Participants were followed for an average of 8.4 years for the men and 7.5 years for the women.

Table 1.1 compares the characteristics of survivors to persons who had died during the follow-up. You can see that there are a number of significant differences between survivors and decedents among men and women. Specifically, survivors were younger, had lower blood pressure, lower cholesterol, were less likely to smoke, and were more physically fit (based on the length of time they stayed on the treadmill and their level of effort).

Although the results are interesting, Table 1.1 does not answer our basic question: Does being physically fit independently increase longevity? It doesn't answer the question because whereas the high-fitness group was less likely to die during the study period, those who were physically fit may just have been younger, been less likely to smoke, or had lower blood pressure.

To determine whether exercise is independently associated with mortality, the authors performed proportional hazards analysis, a type of multivariable analysis. The results are shown in Table 1.2. If you compare the number of deaths per thousand person-years in men, you can see that there were more

[3] Blair, S.N., Kampert, J.B., Kohl, H.W., *et al.* "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women." *JAMA* **276** (1996): 205–10.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

Table 1.1 Baseline characteristics of survivors and decedents, Aerobics Center Longitudinal Study.

	Men		Women	
Characteristics	Survivors (n = 24 740)	Decedents (n = 601)	Survivors (n = 6991)	Decedents (n = 89)
Age, y (SD)	42.7 (9.7)	52.1 (11.4)	42.6 (10.9)	53.3 (11.2)
Body mass index, kg/m^2 (SD)	26.0 (3.6)	26.3 (3.5)	22.6 (3.9)	23.7 (4.5)
Systolic blood pressure, mm Hg (SD)	121.1 (13.5)	130.4 (19.1)	112.6 (14.8)	122.6 (17.3)
Total cholesterol, mg/dL (SD)	213.1 (40.6)	228.9 (45.4)	202.7 (40.5)	228.2 (40.8)
Fasting glucose, mg/dL (SD)	100.4 (16.3)	108.1 (32.0)	94.4 (14.5)	99.9 (25.0)
Fitness, %				
Low	20.1	41.6	18.8	44.9
Moderate	42.0	39.1	40.6	33.7
High	37.9	19.3	40.6	21.3
Current or recent smoker, %	26.3	36.9	18.5	30.3
Family history of coronary heart disease, %	25.4	33.8	25.2	27.0
Abnormal electrocardiogram, %	6.9	26.3	4.8	18.0
Chronic illness, %	18.4	40.3	13.4	20.2

Adapted with permission from Blair, S. N., *et al.* "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women." *JAMA* **276** (1996):205–10. Copyright 1996, American Medical Association. Additional data provided by authors.

deaths in the low-fitness group (38.1) than in the moderate/high fitness group (25.0). This difference is reflected in the elevated relative risk for lower fitness (38.1/25.0 = 1.52). These results are adjusted for all of the other variables listed in the table. This means that low fitness is associated with higher mortality, independent of the effects of other known risk factors for mortality, such as smoking, elevated blood pressure, cholesterol, and family history. A similar pattern is seen for women.

DEFINITION

Stratified analysis assesses the effect of a risk factor on outcome while holding another variable constant.

Was there any way to answer this question without multivariable analysis? One could have performed stratified analysis. Stratified analysis assesses the effect of a risk factor on outcome while holding another variable constant. So, for example, we could compare physically fit to unfit persons separately among smokers and nonsmokers. This would allow us to calculate a relative risk for the impact of fitness on mortality, independent of smoking. This analysis is shown in Table 1.3.

Unlike the multivariable analysis in Table 1.2, the analyses in Table 1.3 are bivariate.[4] We see that the mortality rate is greater among those at low fitness

[4] Some researchers use the term "univariate" to describe the association between two variables. I think it is more informative to restrict the term univariate to analyses of a single variable (e.g., mean, median), while using the term "bivariate" to refer to the association between two variables.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

Table 1.2 Multivariable analysis of risk factors for all-cause mortality, Aerobics Center Longitudinal Study.

	Men		Women	
Independent variable	Deaths per 10 000 person-years	Adjusted relative risk (95% CI)	Deaths per 10 000 person-years	Adjusted relative risk (95% CI)
Fitness				
Low	38.1	1.52 (1.28–1.82)	27.8	2.10 (1.36–3.26)
Moderate/High	25.0	1.0 (ref.)	13.2	1.0 (ref.)
Smoking status				
Current or recent smoker	39.4	1.65 (1.39–1.97)	27.8	1.99 (1.25–3.17)
Past or never smoked	23.9	1.0 (ref.)	14.0	1.0 (ref.)
Systolic blood pressure				
≥140 mm Hg	35.6	1.30 (1.08–1.58)	13.0	0.76 (0.41–1.40)
<140 mm Hg	27.3	1.0 (ref.)	17.1	1.0 (ref.)
Cholesterol				
≥240 mg/dL	35.1	1.34 (1.13–1.59)	18.0	1.09 (0.68–1.74)
<240 mg/dL	26.1	1.0 (ref.)	16.6	1.0 (ref.)
Family history of coronary heart disease				
Yes	29.9	1.07 (0.90–1.29)	12.8	0.70 (0.43–1.16)
No	27.8	1.0 (ref.)	18.2	1.0 (ref.)
Body mass index				
≥27 kg/m^2	28.8	1.02 (0.86–1.22)	15.9	0.94 (0.52–1.69)
<27 kg/m^2	28.2	1.0 (ref.)	16.9	1.0 (ref.)
Fasting glucose				
≥120 mg/dL	34.4	1.24 (0.98–1.56)	29.6	1.79 (0.80–4.00)
<120 mg/dL	27.9	1.0 (ref.)	16.5	1.0 (ref.)
Abnormal electrocardiogram				
Yes	44.4	1.64 (1.34–2.01)	25.3	1.55 (0.87–2.77)
No	27.1	1.0 (ref.)	16.3	1.0 (ref.)
Chronic illness				
Yes	41.2	1.63 (1.37–1.95)	17.5	1.05 (0.61–1.82)
No	25.3	1.0 (ref.)	16.7	1.0 (ref.)

Adapted with permission from Blair, S. N., *et al.* "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women." *JAMA* **276** (1996): 205–10. Copyright 1996, American Medical Association. Additional data provided by authors.

compared to those at moderate/high fitness, both among smokers (48.0 vs. 29.4) and among nonsmokers (44.0 vs. 20.1). This stratified analysis shows that the effect of fitness is independent of smoking status.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

Table 1.3 Stratified analysis of smoking and fitness on all-cause mortality among men, Aerobics Center Longitudinal Study.

	Deaths per 10 000 person-years	Stratum-specific relative risk (95% CI)
Smokers		
Low fitness	48.0	1.63 (1.26–2.13)
Moderate/high fitness	29.4	1.0 (ref.)
Nonsmokers		
Low fitness	44.0	2.19 (1.77–2.70)
Moderate/high fitness	20.1	1.0 (ref.)

Data supplied by Aerobics Center Longitudinal Study.

But what about all of the other variables that might affect the relationship between fitness and longevity? You could certainly stratify for each one individually, proving that the effect of fitness on longevity is independent not only of smoking status, but also independent of elevated cholesterol, elevated blood pressure, and so on. However, this would only prove that the relationship is independent of these variables taken singly.

To stratify by two variables (smoking and cholesterol), you would have to assess the relationship between fitness and mortality in four groups (smokers with high cholesterol; smokers with low cholesterol; nonsmokers with high cholesterol; nonsmokers with low cholesterol). To stratify by three variables (smoking status, cholesterol level, and elevated blood pressure [yes/no]), you would have to assess the relationship between fitness and mortality in eight groups ($2 \times 2 \times 2 = 8$); add elevated glucose (yes/no) and you would have 16 groups ($2 \times 2 \times 2 \times 2 = 16$); add age (in six decades) and you would have 96 groups ($2 \times 2 \times 2 \times 2 \times 6 = 96$); and we haven't even yet taken into account all of the variables in Table 1.1 that are associated with mortality.

With each stratification variable you add, you increase the number of subgroups for which you have to individually assess whether the relationship between fitness and mortality holds. Besides producing mountains of printouts, and requiring a book (rather than a journal article) to report your results, you would likely have an insufficient sample size in some of these subgroups, even if you started with a large sample size. For example, in the Aerobics Center Longitudinal Study there were 25,341 men but only 601 deaths. With 96 subgroups, assuming uniform distributions, you would expect only about six deaths per subgroup. But, in reality, you wouldn't have uniform distributions. Some samples would be very small, and some would have no outcomes at all.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

Multivariable analysis overcomes this limitation. It allows you to simultaneously assess the impact of multiple independent variables on outcome. But there is (always) a cost: The model makes certain assumptions about the nature of the data. These assumptions are sometimes hard to verify. We will take up these issues in Chapters 3, 4, and 9.

1.2 What are confounders and how does multivariable analysis help me to deal with them?

The ability of multivariable analysis to *simultaneously* assess the independent contribution of a number of risk factors to outcome is particularly important when you have "confounding." Confounding occurs when the apparent association between a risk factor and an outcome is affected by the relationship of a third variable to the risk factor and the outcome; the third variable is called a confounder.

DEFINITION

A *confounder* is associated with the risk factor and causally related to the outcome.

For a variable to be a confounder, the variable must be associated with the risk factor and causally related to the outcome (Figure 1.1).

A classically taught example of confounding is the relationship between carrying matches and developing lung cancer (Figure 1.2). Persons who carry matches have a greater chance of developing lung cancer; the confounder is smoking. This example is often used to illustrate confounding because it is easy to grasp that carrying matches cannot possibly cause lung cancer.

Stratified analysis can be used to assess and eliminate confounding. If you stratify by smoking status you will find that carrying matches is not associated with lung cancer. That is, there will be no relationship between carrying matches and lung cancer when you look separately among smokers and nonsmokers (Figure 1.2). The statistical evidence of confounding is the difference

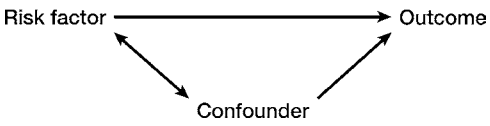

Figure 1.1 Relationships among risk factor, confounder, and outcome.

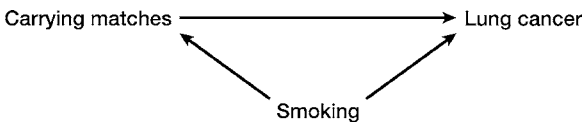

Figure 1.2 Relationships among carrying matches, smoking, and lung cancer.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

between the unstratified and the stratified analysis. In the unstratified analysis the chi-squared test would be significant and the odds ratio for the impact of matches on lung cancer would be significantly greater than one. In the two stratified analyses (smokers and nonsmokers), carrying matches would not be significantly associated with lung cancer; the odds ratio would be one in both strata. This differs from the example of stratified analysis in Table 1.3 where exercise was significantly associated with mortality for both smokers and nonsmokers.

Most clinical examples of confounding are more subtle and harder to diagnose than the case of matches and lung cancer. Let's look at the relationship between smoking and prognosis in patients with coronary artery disease following angioplasty (the opening of clogged coronary vessels with the use of a wire and a balloon).

Everyone knows (although the cigarette companies long claimed ignorance) that smoking increases the risk of death. Countless studies, including the Aerobics Center Longitudinal Study (Table 1.2), have demonstrated that smoking is associated with increased mortality. How then can we explain the results of Hasdai and colleagues?[5] They followed 5437 patients with coronary artery disease, who had angioplasty. They divided their sample into nonsmokers, former smokers (quit at least six months before procedure), recent quitters (quit immediately following the procedure), and persistent smokers. The relative risk of death with the 95 percent confidence intervals are shown in Table 1.4.

How can the risk of death be lower among persons who persistently smoke than those who never smoked? In the case of recent quitters, you would expect their risk of death to return toward normal only after years of not smoking – and even then you wouldn't actually expect quitters to have a lower risk of death than nonsmokers.

Before you assume that there is something wrong with this study, several other studies have found a similar relationship between smoking and better prognosis among patients with coronary artery disease after thrombolytic therapy. This effect has been named the "smoker's paradox."[6] What is behind the paradox? Look at Table 1.5. As you can see, compared to nonsmokers and

[5] Hasdai, D., Garratt, K. N., Grill, D. E., *et al.* "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med.* **336** (1997): 755–61.

[6] Barbash, G. I., Reiner, J., White, H. D., *et al.* "Evaluation of paradoxical beneficial effects of smoking in patients receiving thrombolytic therapy for acute myocardial infarction: Mechanisms of the 'smoker's paradox' from the GUSTO-I trial, with angiographic insights." *J. Am. Coll. Cardiol.* **26** (1995): 1222–9.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

Table 1.4 Bivariate association between smoking status and risk of death.

Bivariate	Nonsmokers	Former smokers	Recent quitters	Persistent smokers
Relative risk of death	1.0 (ref.)	1.08 (0.92–1.26)	0.56 (0.40–0.77)	0.74 (0.59–0.94)

Adapted from Hasdai, D., *et al*. "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med*. **336** (1997): 755–61.

Table 1.5 Association between demographic and clinical factors and smoking status.

	Nonsmokers	Former smokers	Recent quitters	Persistent smokers
Age, year ± SD	67 ± 11	65 ± 10	56 ± 10	55 ± 11
Duration of angina, months ± SD	41 ± 66	51 ± 72	21 ± 46	29 ± 55
Diabetes, %	21%	18%	8%	10%
Hypertension, %	54%	48%	38%	39%
Extent of coronary artery disease, %				
One vessel	50%	51%	57%	55%
Two vessels	36%	36%	34%	36%
Three vessels	14%	13%	10%	9%

Adapted from Hasdai, D., *et al*. "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med*. **336** (1997): 755–61.

former smokers, quitters and persistent smokers are younger, have had angina for a shorter period of time, are less likely to have diabetes and hypertension, and have less severe coronary artery disease (i.e., more one-vessel disease and less three-vessel disease). Given this, it is not so surprising that the recent quitters and persistent smokers have a lower risk of death than nonsmokers and former smokers: They are younger and have fewer underlying medical problems than the nonsmokers and former smokers.

Compare the bivariate (unadjusted) risk of death to the multivariable risk of death (Table 1.6). Note that in the multivariable analysis the researchers adjusted for those differences, such as age and duration of angina, that existed among the four groups.

With statistical adjustment for the baseline differences between the groups, the former smokers and persistent smokers have a significantly greater risk of death than nonsmokers – a much more sensible result. (The recent quitters also have a greater risk of death than the nonsmokers, but the confidence intervals of the relative risk do not exclude one.) The difference between the bivariate and multivariable analysis indicates that confounding is present. The advantage of multivariable analysis over stratified analysis is that it would

TIP

Multivariable analysis is preferable to stratified analysis when you have multiple confounders.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

Table 1.6 Comparison of bivariate and multivariable association between smoking status and risk of death.

	Nonsmokers	Former smokers	Recent quitters	Persistent smokers
Relative risk of death (bivariate)	1.0 (ref.)	1.08 (0.92–1.26)	0.56 (0.40–0.77)	0.74 (0.59–0.94)
Relative risk of death (multivariable)	1.0 (ref.)	1.34 (1.14–1.57)	1.21 (0.87–1.70)	1.76 (1.37–2.26)

Adapted from Hasdai, D., *et al.* "Effect of smoking status on the long-term outcome after successful percutaneous coronary revascularization." *N. Engl. J. Med.* **336** (1997): 755–61.

have been difficult to stratify for age, duration of angina, diabetes, hypertension, and extent of coronary artery disease.

1.3 What are suppressers and how does multivariable analysis help me to deal with them?

TIP

Unlike a typical confounder, when you have a suppresser you won't see any bivariate association between the risk factor and the outcome until you adjust for the suppresser.

Suppresser variables are a type of confounder. As with confounders, a suppresser is associated with the risk factor and the outcome (Figure 1.3). The difference is that on bivariate analysis there is no effect seen between the risk factor and the outcome. But when you adjust for the suppresser, the relationship between the risk factor and the outcome becomes significant.

Identifying and adjusting for suppressers can lead to important findings. For example, it was unknown whether taking antiretroviral treatment would prevent HIV seroconversion among healthcare workers who sustained a needle stick from a patient who was HIV-infected. For several years, healthcare workers who had an exposure were offered zidovudine treatment, but they were told that there was no efficacy data to support its use. A randomized controlled trial was attempted, but it was disbanded because healthcare workers did not wish to be randomized.

Since a randomized controlled trial was not possible, a case-control study was performed instead.[7] The cases were healthcare workers who sustained a needle stick and had seroconverted. The controls were healthcare workers who sustained a needle stick but had remained HIV-negative. The question was whether the proportion of persons taking zidovudine would be lower in the group who had seroconverted (the cases) than in the group who had not become infected (the controls). The investigators found that the proportion of cases using zidovudine was lower (9 of 33 cases or 27 percent) than the

[7] Cardo, D.M., Culver, D.H., Ciesielski, C.A., *et al.* "A case-control study of HIV seroconversion in health-care workers after percutaneous exposure." *N. Engl. J. Med.* **337** (1997): 1485–90.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-14107-9 - Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers, Third Edition
Mitchell H. Katz
Excerpt
More information

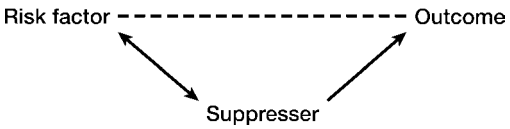

Figure 1.3 Relationships among risk factor, suppresser, and outcome.

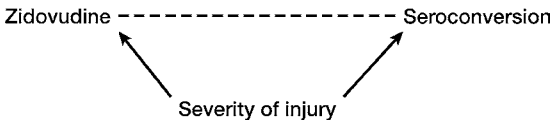

Figure 1.4 Relationships among zidovudine, severity of injury, and seroconversion.

proportion of controls using zidovudine (247 of 679 controls or 36 percent), but the difference was not statistically significant (probability [*P*] = 0.35). Consistent with this nonsignificant trend, the odds ratio shows that zidovudine was protective (0.7), but the 95 percent confidence intervals were wide and did not exclude one (0.3–1.4).

However, it was known that healthcare workers who sustained an especially serious exposure (e.g., a deep injury or who stuck themselves with a needle that had visible blood on it) were more likely to choose to take zidovudine than healthcare workers who had more minor exposures. Also, healthcare workers who had serious exposures were more likely to seroconvert.

When the researchers adjusted their analysis for severity of injury using multiple logistic regression, zidovudine use was associated with a significantly lower risk of seroconversion (odds ratio [OR] = 0.2; 95 percent confidence interval (CI) = 0.1 – 0.6; $P < 0.01$). Thus, we have an example of a suppresser effect as shown in Figure 1.4. Severity of exposure is associated with zidovudine use and causally related to seroconversion. Zidovudine use is not associated with seroconversion in bivariate analyses but becomes significant when you adjust for severity of injury.

Although this multivariable analysis demonstrated the efficacy of zidovudine on preventing seroconversion by incorporating the suppresser variables, it should be remembered that multivariable analysis cannot adjust for other potential biases in the analysis. For example, the cases and controls for this study were not chosen from the same population, raising the possibility that selection bias may have influenced the results. Nonetheless, on the strength of this study, postexposure prophylaxis with antiretroviral treatment became the standard of care for healthcare workers who sustained needle sticks from HIV-contaminated needles.

© in this web service Cambridge University Press www.cambridge.org