

## THE ROLE OF MODEL SELECTION IN CAUSAL INFERENCE FROM NONEXPERIMENTAL DATA

JAMES M. ROBINS<sup>1</sup> AND SANDER GREENLAND<sup>2</sup>

The article by Starr et al. (1) in this issue of the *Journal* provides a valuable starting point to examine the role of model selection when using multivariate models in causal inference. In their discussion, Starr et al. make two observations that taken together raise a central problem. They first note that "the complementary nature of the relationship between some explanatory variables can lead to compromised inferences if one or the other is innocently omitted from consideration." Epidemiologists should recognize this as the familiar problem of confounding. But Starr et al. go on to note that "the indiscriminate inclusion of any and all variables that might just possibly be important" may also compromise inferences. They thus imply (and we agree) that inferences regarding an exposure effect may be compromised by the inclusion of too few or too many covariates in a model. Faced with this dilemma, the authors adopt a backward-elimination model-selection strategy, allowing for ad hoc modifications such as forcing in certain main effects believed a priori to be nonzero. While we do not take issue with the need for some sort of model-selection strategy such as Starr et al. and others (2) have employed, we feel that the descriptions and justifications offered for such strategies have been incomplete. We will argue that all modelling strategies contain implicit prior beliefs about nature. One can justify a modelling strategy only if the

prior beliefs implicit in the strategy reflect one's actual prior beliefs. As such, modelling strategies should be viewed as attempts to approximate a formal Bayesian analysis (in which one first quantifies one's prior beliefs using a subjective probability distribution, then uses the data to update this distribution by Bayes' rule (3)). Many of our observations echo more complete arguments found in the statistical literature (3, 4).

This paper will focus on the use of models for the control of multiple covariates when estimating exposure effects from nonexperimental data. We will not consider other common uses of models, such as providing parsimonious summaries of complex data sets or testing theories of causal mechanism (as, for example, when an investigator examines the fit of a multi-stage cancer model).

For the sake of simplicity, we will ignore problems of misclassification and subject selection, and we will for the most part ignore the distinction between small-sample (exact) and large-sample (asymptotic) properties, although our comments apply when such considerations are taken into account. All our references to means, variances, standard errors, and confidence intervals will implicitly refer to the large-sample versions of these quantities. We will assume that each study subject is randomly sampled from a very large target population (or hypothetical superpopulation), and the sampling variability of estimates refers to hypothetical resampling of the target population. (Note that if we assumed that each person's outcome was completely predetermined, then such a sampling model is necessary to give meaning to variances and

<sup>1</sup> Occupational Health Program, Harvard School of Public Health, 665 Huntington Ave., Boston, MA 02115. (Send requests for reprints to Dr. James M. Robins at this address.)

<sup>2</sup> Division of Epidemiology, UCLA School of Public Health, Los Angeles, CA.

The authors thank Hal Morgenstern for his helpful comments on the manuscript.

confidence intervals from a nonexperimental study.) Finally, we will assume that the primary goal in the analysis is the estimation of the effect of a specific study exposure on risk of a specific study disease in the target population.

## THE ROLE OF MODELS IN CAUSAL INFERENCE

### *Preliminary concepts*

Epidemiologists usually quantify effects in terms of either relative risks or regression coefficients. In this section, we will quantify the effect of an exposure in a given exposed population (or stratum thereof) in terms of the standardized morbidity ratio parameter, or SMR (5), defined as the ratio of the number of cases that should be expected over the study period (given that the population is exposed) to the number that would have been expected had the same population been unexposed.

In this paper, we will say an estimator of an effect measure is unbiased if one can obtain valid (large-sample) 95 per cent confidence intervals for the measure by taking the estimator plus or minus 1.96 standard errors (this property has been termed *uniform asymptotic unbiasedness* (6)). This definition corresponds closely to the intuitive notions of unbiasedness used in epidemiologic discussions of confounding (5-9). In repetitions of the study, biased estimators will systematically deviate from the true value of the effect measure, and therefore cannot be used to construct valid confidence intervals. The (large-sample) mean value of this deviation we will call the *bias* of the estimator. We will call a covariate a "confounder" if estimators which are not adjusted for the covariate are biased. Following common usage, we will say an estimator is "confounded" if it is not adjusted for a confounder. As an example, consider a cohort study of the effect of asbestos exposure on lung cancer risk, and suppose the exposed and unexposed groups were at start of follow-up similar in all respects except for smoking habits. If smoking were

positively associated with asbestos exposure, we would expect that, upon repetitions of the study, the crude risk ratio estimator would systematically overestimate the true standardized morbidity ratio parameter, and 95 per cent confidence intervals based on the crude estimator would contain the true parameter less than 95 per cent of the time. In other words, the crude risk ratio would be confounded by smoking.

Finally, a *statistical model* is a mathematical expression for a set of assumed restrictions on the possible states of nature. For example, a linear (i.e., main-effects only) logistic model for the dependence of subsequent fertility on dibromochloropropane exposure and parity implies the following restrictions about nature: 1) an exponential dependence of the fertility odds on dibromochloropropane and parity; 2) a constant odds ratio across dibromochloropropane for the association of any parity level (relative to zero parity) with subsequent fertility; and 3) a constant odds ratio across parity for the association of any dibromochloropropane level (relative to zero exposure) with subsequent fertility.

### *Why do we need models to estimate effects?*

By measuring more and more potential confounders, one can increase one's subjective belief that there is little residual confounding by unmeasured covariates. But one then faces the analytic problem of estimating effects while controlling for a large number of variables. It is here that one would naturally consider using multivariate models.

To examine the analytic problem further, consider an unmatched cohort study in which data on a dichotomous exposure and on 20 dichotomous covariates (believed to be potential confounders) have been obtained on 1,000 study subjects. Suppose that after stratifying on all 20 covariates no residual confounding exists by unmeasured risk factors. For each possible combination of values for the 20 covariates, one can construct a (stratum-specific)  $2 \times 2$  table

of exposure level and disease outcome among stratum members; there will be  $2^{20}$  (or about a million) such tables. For each stratum, there are two parameters that determine the outcomes in that stratum: an effect parameter (which we have taken as the risk ratio) and a "nuisance" parameter (which we would take as the risk among the unexposed, i.e., the baseline risk). Thus there are a total of  $2 \times 2^{20}$  unknown parameters.

Unfortunately, with the 1,000:1 ratio of strata to subjects in this example, one should expect to find few (if any) strata that contained both an exposed subject and an unexposed subject. Suppose (as would be likely) no stratum contained more than a few subjects, and only a few strata contained both exposed and unexposed subjects. Because of the frequency of zero denominators, without making any assumptions one could learn very little about the effect of exposure in any stratum, or about the standardized morbidity ratio parameter. This reflects the fact that stratum level and exposure status are almost perfectly associated in the data; thus the effects of exposure cannot be separated from the effects of the covariates, and one must seek "help" from statistical models.

In our example, all the possible states of nature are described by all the possible combinations of values for the  $2^{20}$  unknown risk ratios and the  $2^{20}$  baseline risks. Even if one makes no assumptions about nature, the resulting view may be termed a "model." This "model" of nature (incorporating zero restrictions) is called the "saturated model," for it is a model that is "saturated" with unknown parameters.

One model restriction the investigator might make on nature would be to assume that the risk ratio is constant across strata, which we shall call "Model I"; in our example, this model has  $2^{20} + 1$  unknown parameters. If Model I were in fact correct, any unbiased estimator of the constant risk ratio (e.g., the Mantel-Haenszel risk ratio (2, 10)) would be an unbiased estimator of the true standardized morbidity ratio pa-

rameter; it would also have uselessly large variance, in that confidence limits constructed from the estimator would be too wide to tell us anything interesting about the parameter. Thus, assuming Model I would not solve our problem.

Another way to view the problem with Model I is as follows: one wants to report an estimate of the exposure effect that is close on average (i.e., in hypothetical repetitions) to the true but unknown value. A common measure of closeness is the average of the squared distance between an estimate and the true parameter—the mean-squared error. The mean-squared error is equal to the variance of the estimator plus the square of the bias. Any unbiased estimator based on Model I (i.e., stratifying on all 20 covariates) has too large a variance and thus too large a mean-squared error to be useful.

As a result of the problem with Model I, one might entertain a new model, which we will call Model II, in which not only is the risk ratio assumed to be constant over strata but covariates 5–20 are assumed to have no predictive value for disease outcome upon controlling for exposure and covariates 1–4, i.e., only covariates 1–4 are considered independent risk factors. This model has  $2^4 = 16$  "nuisance" parameters and one parameter of interest, for a total of only 17 unknown parameters. If Model II is true of nature, the Mantel-Haenszel (and maximum-likelihood, etc.) estimator under Model II will be an unbiased estimator of the exposure effect, and will in general have much smaller variance than any estimator unbiased under Model I. This is because using Model II is equivalent to stratifying on covariates 1–4 only; this leads to only 16 strata in the analysis, and with 1,000 subjects one could expect at least a few of them to have many exposed and unexposed subjects. We would informally interpret the improvement in precision going from Model I to Model II as follows: whatever model is used, each parameter specified frees up some of the information in the data for estimation of the remaining

parameters. In using Model II instead of Model I, only 2<sup>4</sup> nuisance parameters are left unspecified, and thus more data information is left for estimation of the parameter of interest.

But what if one of covariates 5-20, say covariate 6, is in fact a risk factor and is associated with exposure? Then in general an estimator based on Model II (i.e., controlling only covariates 1-4) will be a biased estimator of the exposure effect, as it will be confounded by covariate 6. Nevertheless, the variance of a Model II estimator will still be less than that of an estimator based on Model I. Consequently, the mean-squared error of a Model II estimator could be either greater or less than that of a Model I estimator, depending on the bias of the Model II estimator. For example, it would have a smaller mean-squared error if Model II is nearly but not quite correct (because its bias would then be small), but it might have larger mean-squared error if Model I is nearly correct but Model II is grossly incorrect (because its bias would then be large).

The point of the above example is this: the only legitimate justification for choosing a particular model for analysis is that one believes the savings in variance afforded by making the model assumptions about nature will offset the increase in bias that will result from those assumptions being incorrect. Since the true state of nature is unknown, the magnitude of the bias cannot be known. Thus, one's choice is subjective and subject to error. This choice will be most difficult when, as in our example, one is confronted with "weak data", that is, data such that no estimator believed with near certainty to be unbiased for the parameter of interest has a variance small enough to allow construction of useful confidence intervals. With weak data, we cannot make accurate inferences about the parameter of interest unless we incorporate into the analysis substantially correct assumptions about nature. (In the above example, had we been able to match jointly on most of the 20 covariates, we would have

had sparse but not weak data. Effect parameters can often be efficiently and unbiasedly estimated from sparse data, although special estimators need to be used (10, 11).)

#### INTERPRETATION OF MODEL-SELECTION STRATEGIES

We have pointed out that an estimator of exposure effect based on a highly-saturated model (i.e., a model that includes terms for many or most covariate effects and interactions) will have small bias provided that no confounding by unmeasured covariates remains, but may have large variance. On the other hand, a model with only a few covariates and interactions (i.e., a "reduced" model) is likely to be grossly inconsistent with the true state of nature (i.e., highly misspecified) and thus may yield a very biased (though less variable) estimator of exposure effect. Thus, the mean-squared error of the estimator from either a saturated or reduced model may be quite large, and one might wish to consider using instead an estimator derived under a model-selection strategy such as backward elimination (a stepwise regression procedure).

An estimator is a rule that, when applied to a data set, produces an estimate of the parameter of interest. The backward-elimination estimator is given by the following rule: start with a highly-saturated model and eliminate terms one by one until only "significant" and "forced-in" terms are left, then use the effect estimate from the resulting model. The backward-elimination estimator is in general a biased estimator (6), and will have a smaller mean-squared error than the saturated and reduced model estimators only for certain states of nature (i.e., only at certain points in the parameter space of the saturated model).

Since the choice of the best estimator depends on the unknown state of nature, one can only choose among estimators based on one's beliefs about what the true state of nature is. One can justify a model-selection strategy based on significance

tests of coefficients (such as the backward-elimination strategy described by Starr et al. (1) and others (2)) only if one believes that 1) the covariates forced in the model have dose-response relationships close to those implied by the model, and 2) the coefficients of covariates subject to elimination are probably near zero but may (with small subjective probability) be very different from zero. To see why this is so, note that if one believed with absolute certainty that a particular covariate's coefficient was zero, one should attribute the data evidence that the coefficient is nonzero to sampling variation and therefore set that coefficient to zero (no matter how small the  $p$  value found in a test of that coefficient in the starting model). A test of the coefficient is thus unnecessary, and the "reduced" model that mandatorily leaves the variable out will yield an effect estimator with a mean-squared error less than that of the backward-elimination estimator (provided the covariate's coefficient actually is zero). As an example, if Starr et al. had data on city of birth, their decision not to allow this covariate to enter any model would presumably represent a near-certain prior belief that its coefficient was close to zero.

In practice, one should want to protect oneself from incurring a large mean-squared error if one's prior beliefs turn out to be far from correct. Thus, if one's beliefs are badly contradicted by the data, one should be willing to give them up, and the  $\alpha$  level chosen for the coefficient test should depend in part on the strength of one's beliefs. For example, with strong but not certain belief that a particular coefficient is near zero, a test with a 0.05  $\alpha$  level might be appropriate.

If one accepts the preceding arguments, one must recognize that zero often has no special claim as the value to test against. If, for example, Starr et al. had believed a priori that the most likely value of the coefficient for their covariate lag<sub>1</sub> was  $-0.2$ , it might make sense to test whether the coefficient for lag<sub>1</sub> was significantly differ-

ent from  $-0.2$  and set the coefficient to  $-0.2$  if the test fails to reject. In the absence of prior beliefs, some investigators treat zero as the most likely value (and thus test against it) based on the principle of parsimony. Nevertheless, any parsimony principle is irrelevant for a coefficient that one believes a priori is nonzero.

Even if one believed that the most likely value for a covariate's coefficient was zero, why should one use a decision rule that uses the data estimate from an expanded model if the test rejects zero, but otherwise uses one's prior best guess? If the estimated covariate coefficient was (say)  $-0.3$  with a standard error of 0.3, some value of this coefficient between  $-0.3$  and zero might be one's revised best guess for the value of the coefficient after seeing the data. As a first approximation, weighted averages of one's prior beliefs and data information would often seem preferable to us, although proper combination of prior information with data information involves more complex procedures (cf., Section 5.8 of reference 3).

If one is sure a covariate has a nonzero coefficient, we suspect that strategies that force the covariate into the regression do not go far enough. Consider, for example, Starr et al.'s situation: what if the parity estimate was 0.0 with a confidence interval of  $(-1.0, 1.0)$ ? In such a case, forcing parity in would result in the same point estimate of the exposure effect as leaving parity out. Again, it seems to us that (as a first approximation) one would be better off setting the coefficient for parity to some number (greater than 0.0) representing a weighted average of one's prior beliefs and the data information.

#### *Confidence intervals and confounding*

Thus far we have only addressed the mean-squared error of point estimators of the exposure effect. How should one assess their precisions? Even when there is no confounding by unmeasured covariates, confidence intervals (and  $p$  values) obtained after backward elimination are in-

valid, in that there will be states of nature (i.e., points in the parameter space of the saturated model) for which the intervals will fail to cover the effect parameter of interest at the nominal coverage rate (6). The coverage can be much less than nominal when the power to detect a nonzero coefficient is low. In typical data sets (such as Starr et al.'s), procedures based on significance tests of 0.05 level would be quite likely to set to zero some coefficients whose true value was large.

Note that the use of significance tests in model selection when estimating causal parameters goes against recommendations in the epidemiologic literature that one should not test the association of a confounder with disease (2, 5, 7, 8). Robins and Morgenstern (9) have interpreted these recommendations to mean that the epidemiologic concept of confounding is intimately associated with the idea of valid confidence intervals. To obtain valid confidence intervals in the absence of perfectly correct prior beliefs, one must use a completely saturated model. But as we have noted, data containing measurements on many potential confounders will usually be weak, and will yield prohibitively wide saturated-model intervals. Thus, to obtain useful intervals from such data, one will usually end up setting some nonzero coefficients to zero and hence introduce bias and sacrifice confidence interval validity (i.e., confounding will be introduced). In general, as one allows the  $\alpha$  level of the significance tests used in a backward-elimination procedure to increase, the difference between the nominal and true coverage rates of one's "confidence intervals" will decrease, but, unfortunately, with weak data the width of the intervals will usually increase.

Although a confidence interval derived after variable selection will in general be invalid, it may sometimes serve as a rough measure of our posterior uncertainty and could then be used to construct an informal subjective posterior distribution for the effect parameter. We discuss this in more detail below.

#### FURTHER ANALYSIS OF MODEL-SELECTION STRATEGIES

The arguments given above would ultimately lead one to do a formal Bayesian analysis in which one would write down one's joint subjective probability distributions for all the unknown parameters; the data (more precisely, the likelihood function of the saturated model) would then be used to update these prior distributions by Bayes' rule (3). The optimal point estimate for the exposure effect (in terms of mean-squared error) would simply be the mean of the posterior distribution of the effect parameter. Unfortunately, formal multivariate Bayesian analysis is nearly impossible to carry out due to the difficulties of eliciting multivariate prior distributions and computing posterior distributions. Thus, in practice, one is led to employ informal approximations to a Bayesian analysis, such as the ones described above. Since evaluating the accuracy of such approximations would require actually carrying out the Bayesian analysis, and since such evaluations have not been done, tests and interval estimates resulting after variable selection do not possess a formal Bayesian justification (just as they do not possess their nominal frequentist interpretation). Nevertheless, as we have shown above, one can do an analysis of their behavior under hypothetical repetitions of the study; one can also attempt to infer the prior beliefs that underlie the analysis choices made by investigators.

#### *Three backward-elimination strategies*

We have documented the set of prior beliefs that justify the use of the backward-elimination strategies used by Starr et al. (1) and others (2), but we have acted as if these strategies represent a well-defined unique method. In practice, however, we have observed at least three different backward-elimination strategies for estimating an exposure effect: 1) eliminate terms until only significant terms remain in the model—if the study exposure is eliminated,

infer the exposure has no effect, otherwise use the estimate of (and confidence interval for) the exposure effect from the final model; 2) eliminate terms until either only significant terms remain or the study exposure is next to remove, then use the final model estimate of (and confidence interval for) the exposure effect; and 3) eliminate terms until only significant terms remain, except keep exposure in the model and use the final model estimate of (and confidence interval for) the exposure effect. The third method is the method that appears closest to that described by Starr et al.; unfortunately, this method can also be highly biased.

Consider a study in 1930 (before the hypothesis that cigarette smoking caused lung cancer was established as fact) in which match carrying was the study exposure, cigarette smoking the potential confounder, and lung cancer the study disease. Suppose also that the amount of data was limited. For most data realizations, match carrying and smoking would be so correlated that, in a model including both, neither would be significant after adjusting for the other. This result appropriately reflects the fact that it would be impossible to determine from the data alone whether the observed variation in lung cancer risk was due to smoking, to match carrying, or to both. Since the real effects would be due to smoking, the  $p$  value for smoking effect would usually be smaller than that for match carrying. Nevertheless, using the third backward-elimination method, smoking would usually be eliminated from the model, because smoking would usually be nonsignificant in the starting model. With the deletion of smoking, match carrying would become "highly significant" because it would be a proxy for smoking. Thus, the third method would be almost certain to determine that match carrying is an important risk factor with a narrow "confidence interval." The only justification we can see for using the third method is if one has strong prior beliefs both that exposure has an effect, and that the effect of any correlated covariate is small. These are not the

prior beliefs one would have in most studies.

Backward-elimination methods 1 and 2 do not fare quite as poorly in the preceding example. In most data realizations, method 2 would result in both match carrying and smoking remaining in the model; occasionally, smoking would be eliminated and match carrying would appear to be a strong risk factor with a narrow "confidence interval." In most data realizations, method 1 would show smoking to be a strong risk factor with a narrow confidence interval (which would be an *inappropriate* interpretation of the data unless one had strong prior beliefs that match carrying has no effect), although occasionally match carrying would appear to be a strong risk factor with a narrow confidence interval (which would also be an inappropriate interpretation unless one had strong prior beliefs that smoking has no effect). Note that method 1 also encourages confusion of "lack of statistical significance" with "no effect."

We wish to emphasize that in absence of prior knowledge, a model that always included both smoking and match carrying would be the only appropriate analysis. Of course, given our present knowledge of lung cancer biology, we know that we should never enter match carrying in the analysis, since we are certain that its coefficient is virtually zero.

#### *Algorithms versus practice: an informal Bayesian perspective*

A closer scrutiny of Starr et al.'s discussion indicates that, in drawing their final conclusions, they would not rigidly adhere to any mechanical variable-selection strategy. Starr et al. note that both parity and the study exposure (dibromochloropropane) are known a priori to be important determinants of subsequent fertility. They consequently use the nonsignificance of parity in the reduced model for plant A as evidence against the validity of that model, and the nonsignificance of dibromochloropropane in a highly-saturated model as evidence against that model. We interpret their comments as indicating that Starr et

al.'s implicit variable-selection strategy is to choose a model that is consistent with the data and yields parameter estimates consistent with their prior beliefs. Regardless of whether our interpretation of their behavior is correct, this strategy has an informal Bayesian justification that we now illustrate.

*Example.* Suppose that Starr et al.'s backward-elimination strategy resulted in a model in which dibromochloropropane significantly increased fertility with a "confidence interval" for the exposure coefficient of (0.2, 1.0). This result is contrary to our prior beliefs (from animal and human studies) that exposure is supposed to reduce fertility. In such a situation, it is likely that the "confidence interval" from Starr's saturated model would be much wider; in particular, the lower limit derived from the saturated model would likely be less in conflict with our prior beliefs, even if the point estimate was unchanged. For concreteness, suppose that the saturated-model interval was (-0.6, 1.8). One might in this case be inclined to choose the saturated-model interval as one's estimate, rather than the backward-elimination interval, precisely because of the sharp conflict of the latter with one's prior belief, and despite the fact that one would have chosen the backward-elimination interval had it not so sharply conflicted with one's prior belief. The penalty incurred for this choice is a much wider interval.

An informal Bayesian explanation for this penalty is as follows: If the exposure estimate agrees with one's prior beliefs (as in the paper by Starr et al.), the variance of our posterior subjective distribution for exposure effect will be small, as the data have increased one's certainty about the size of the effect; this increased certainty is reflected by the narrow backward-elimination interval reported by Starr et al. But if the exposure estimate differs from one's prior expectation, as in our example, the standard error of one's posterior distribution can be large, as one's uncertainty about the exposure effect may have been increased by the discrepancy; this increased

uncertainty is reflected in the wide saturated-model interval. (In light of our prior beliefs concerning the DBCP effect, an interval whose upper and lower limits are somewhat less than the corresponding upper and lower limits of the saturated model interval might be preferable to the saturated model interval as an approximation to our posterior uncertainty.)

In taking a Bayesian view, one must consider (in addition to the exposure estimate) the plausibility of the model with respect to all the covariates for which one has strong prior beliefs. Any conflicts should contribute to one's uncertainty regarding the correctness of one's data and prior beliefs; in the above example, this forces one back toward the wider intervals of more saturated models.

#### ADDITIONAL CONSIDERATIONS

##### *The goodness-of-fit criterion*

"Goodness-of-fit" is sometimes used as a justification for estimating exposure effects from a particular model (where "goodness-of-fit" is some criterion—such as a likelihood-ratio test—that takes account of both the number of model parameters and ability of the model to reproduce the data). For example, in one backward-elimination method, the computer throws terms out of the model (i.e., sets coefficients of successive covariates to zero) until throwing out any additional covariates would produce a "significant" goodness-of-fit test. The problem with the goodness-of-fit criterion is that, in general, with weak multivariate data: 1) the number of "good fitting" models will be large; 2) there can be large variation in the magnitude of the estimated exposure among the models with good fit; and 3) the intermodel variation in estimated exposure effect can greatly exceed the model-specific standard errors for the exposure effect spewed out by the computer. (An example of all three points would be given by the smoking-match-carrying example discussed earlier.) Thus, one must either choose among the numerous "good fitting" models, on the basis of one's subjective

beliefs, or use the estimate and standard error from the saturated model when the uncertainty associated with the saturated model estimate more closely represents one's subjective degree of uncertainty.

Another problem with goodness-of-fit tests is that they are insensitive to certain types of inconsistencies between models and data, and so they can indicate a good fit for certain models that are grossly inconsistent with the data. Thus, if one wishes to use a model consistent with the data, one ought to check for model adequacy by examining residuals, screening for outliers, and employing other similar "diagnostic techniques." But even among the class of models consistent with the data, the intermodel variation in the estimated exposure effect may exceed the model-specific "standard errors."

#### *Modelling the exposure-covariate associations*

Thus far, we have only considered the use of prior beliefs concerning the strength of a covariate as an independent risk factor for disease (i.e., beliefs concerning the population covariate-disease association conditional on exposure status and the levels of other covariates). Since the exposure-covariate association (in the source population) is also an important determinant of confounding in case-control studies, one might also wish to take account of prior beliefs concerning the population exposure-covariate association in case-control analyses; this would involve the modelling of the exposure-covariate associations among the controls. Unlike relative-risk regression models (such as logistic models), contingency-table models (such as log-linear models) allow simultaneous modelling of both the exposure-covariate and covariate-disease associations. The previous lack of enthusiasm for modelling the population exposure-covariate associations may reflect the fact that our prior beliefs concerning these associations are usually less sharp than those concerning the disease-covariate associations. (Rosenbaum and Rubin (12) have proposed modelling of the mar-

ginal exposure-covariate associations in follow-up studies. Although their approach may be of practical value, the authors' theoretical justification conflicts with the Bayesian and conditional frequentist points of view (6, 13); epidemiologists have also raised theoretical objections to such modelling (5, 7, 8).)

#### *Model-selection in randomized studies*

In a randomized trial, it is not necessary to collect data on covariates in order to obtain valid confidence intervals for the overall treatment effect when, for example, we measure the overall treatment effect by the ratio of the number of cases expected had the entire study population been treated to the number expected had it remained untreated. If, however, data on a risk factor are collected and there is a chance treatment-factor association in the data, from a Bayesian or conditional frequentist point of view one should consider the factor a confounder and adjust for it in the analysis (6). This point has also been made in the epidemiologic literature (5, 7, 8). (A conditional frequentist examines the crude estimator only over those hypothetical repetitions of the trial in which the measured risk factors have exactly the same associations with treatment and each other as was observed; thus, under this conditional view, the crude estimator of effect will in general be biased if any measured risk factor is associated with treatment. It follows from our definition of a confounder that a measured risk factor associated with treatment is a confounder (6).) In other words, although Bayesians and conditional frequentists may find randomization useful in guarding against confounding by unmeasured covariates, in terms of controlling confounding by measured covariates, it is irrelevant whether the data were obtained from a nonrandomized or randomized study (5-9, 13). As such, one might naively expect that the danger of indiscriminately using a "canned" method of backward elimination (such as method 3 described above) to adjust for measured covariates would be as great with random-

ized as with nonrandomized studies. This is not so, because setting a nonzero coefficient to zero will not lead to a significant bias in the estimate of the overall treatment effect unless a strong treatment-covariate association exists in the data, and, in a large trial, randomization makes it improbable that a strong treatment-covariate association exists.

Even when data on a huge number of covariates are collected (so that by chance some covariates are almost certain to be strongly associated with exposure), randomization makes it highly probable that the set of risk factors whose coefficients were incorrectly set to zero by the backward-elimination procedure will contain similar proportions of positive confounders and negative confounders. Furthermore, if only a moderate number of strong risk factors are measured, it is highly improbable that the measured factors strongly associated with exposure will also be those that are the strong risk factors. Thus, in most (but not all) large randomized trials, little bias in the estimator of the overall treatment effect should ensue from using backward-elimination techniques.

#### RECOMMENDATIONS AND CAUTIONS

In nonexperimental studies, one often collects data on many potential confounders. Even in the absence of confounding by unmeasured risk factors, the resulting data will be weak data (unless the sample size is very large or matching was done on most of the confounders). As Starr et al. note, one then faces the dilemma that both improper omission and indiscriminate inclusion of variables in a model can lead to compromised inferences. To explore this issue further, we have dissected one family of variable selection strategies (backward-elimination methods) to show that, from a frequentist perspective, such strategies (including the one described by Starr et al.) can easily lead to highly biased estimators and almost inevitably lead to invalid confidence intervals and *p* values. Nevertheless, we must recognize that with weak data we have little choice but to incorporate

prior beliefs into the analysis (and thus introduce bias and confounding). We thus would operationalize Starr et al.'s recommendation for "considerable care and thought" by 1) writing down in advance our prior beliefs about the effects of candidate variables, 2) not employing a "canned" model selection strategy that implies prior beliefs in serious conflict with our actual prior beliefs, and 3) not accepting results from a model that forces or estimates any parameter to have a value in serious conflict with strong prior beliefs (even if the model fits well).

Unfortunately, following these recommendations will not result in foolproof inferences. For example, suppose 1) the assumptions about nature implicit in one's model are sharp (for example, we select a model with few parameters) and reflect one's actual prior beliefs, 2) the data are weak, 3) the parameter estimates given by the model are consistent with one's prior beliefs, *but* 4) the assumptions implicit in the model differ markedly from the true state of nature. Then our inferences ("confidence intervals") will appear sharp but can be quite wrong, *even if we are careful to choose models that provide a good fit to the data* (4). To quote David Freedman (14) paraphrasing Freedman et al. (15): "It ain't what you don't know that gets you into trouble, it's what you think you know that ain't so. Statistical modelling seems likely to increase the stock of things you think you know but ain't so."

Nevertheless, studies are often used in making decisions (even if the decision is only whether to conduct further research on the exposure under study). If one must make a decision based in part on weak data, one must use one's (possibly incorrect) prior beliefs. A prior belief should reflect one's actual beliefs about a parameter; but, when one uses a "canned" variable-selection algorithm (such as found in typical stepwise regression programs), one is in effect using sharp, arbitrary "priors" and so one is likely to get wrong inferences out of such programs (usually via the improper deletion of important confounders (8)).

The above cautions lead us to recommend that one always perform sensitivity analyses. That is, one should determine if moderate changes in one's prior beliefs or one's analytic procedures (such as one's model-selection strategy) would lead to large changes in one's inferences about the parameter of interest. With weak nonexperimental data, strong dependencies of inferences on prior beliefs will usually be found.

## DISCUSSION

We have argued that our inability to estimate effects accurately from nonexperimental data is not resolved by simply measuring more and more potential risk factors until eventually all important risk factors have been measured. If many strong risk factors are associated with exposure, one still needs to have nearly correct prior beliefs concerning not only which of the potential risk factors are truly important risk factors, but also the magnitude of their effects.

Viewed in this light, the central issue in the choice between indirect standardization (as employed by Starr et al. (1)) and modelling is whether one subjectively believes that US population rates are appropriate to use as prior information about the study cohort. But this choice is really not one between standardization and modelling, because (as Starr et al. seem to be aware) information about US population rates can be incorporated into multivariate modelling; methods for doing so are discussed in detail by Breslow et al. (16). As Starr et al. note in their closing paragraph, indirect standardization involves no assumptions about linearity of dose-response or the order of interactions between the covariates. Such assumptions are replaced by the assumption that US population rates reflect what the study population would have experienced had it not been exposed. Starr et al. fail to mention, however, that comparisons of different standardized fertility ratios or standardized morbidity ratios obtained from indirect standardization

may be invalid unless the interactions of the exposure effect with the covariate effects are perfectly multiplicative, i.e., the covariate-specific fertility ratios must be constant (16, sec. 5.1). Thus, valid use of indirect standardization is not entirely free of modelling assumptions. Nevertheless, we agree with Starr et al. that, in the absence of a sound biologic basis for choosing a particular model, some form of standardization should continue to be employed as part of routine analysis.

## REFERENCES

1. Starr TB, DalCorso RD, Levine RJ. Fertility of workers: a comparison of logistic regression and indirect standardization. *Am J Epidemiol* 1986;123:490-8.
2. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods*. Belmont, CA: Lifetime Learning Publications, 1982.
3. Leamer E. *Specification searches*. New York: John Wiley & Sons, 1978.
4. Dempster AP. Purposes and limitations of data analysis. In: Box GE, Leonard C-F-W, eds. *Scientific Inference: Data Analysis and Robustness (Symposium)*. New York: Academic Press, 1983.
5. Miettinen OS, Cook EF. Confounding: essence and detection. *Am J Epidemiol* 1981;114:563-603.
6. Robins JM. *The statistical foundations of confounding in epidemiology*. Technical report no 2. Boston, MA: Occupational Medicine Program, Harvard School of Public Health, 1983.
7. Rothman KJ. *Epidemiologic methods in clinical trials*. *Cancer* 1977;39:1771-5.
8. Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. *Int J Epidemiol* 1980;9:361-7.
9. Robins JM, Morgenstern H. *Confounding and prior knowledge*. Technical report no 1. Boston, MA: Occupational Medicine Program, Harvard School of Public Health, 1983.
10. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 1985;41:55-68.
11. Breslow NE. Odds ratio estimators when the data are sparse. *Biometrika* 1981;68:73-84.
12. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984;79:516-24.
13. Basu D. Rejoinder. *J Am Stat Assoc* 1980;75:593-5.
14. Freedman DA, Navidi WC. *Regression models for adjusting the 1980 Census*. Technical report no 35. Berkeley, CA: Department of Statistics, University of California, 1984.
15. Freedman DA, Rothenberg T, Sutch R. On energy policy models. *J Bus Econ Stat* 1983;1:24-36.
16. Breslow NE, Lubin JH, Marek P, et al. Multiplicative models and cohort analysis. *J Am Stat Assoc* 1983;78:1-12.