Invited Commentary

Invited Commentary: Variable Selection versus Shrinkage in the Control of Multiple Confounders

Sander Greenland

From the Departments of Epidemiology and Statistics, University of California, Los Angeles, CA.

Received for publication February 5, 2007; accepted for publication September 12, 2007.

After screening out inappropriate or doubtful covariates on the basis of background knowledge, one may still be left with many potential confounders. It is then tempting to use statistical variable-selection methods to reduce the number used for adjustment. Nonetheless, there is no agreement on how selection should be conducted, and it is well known that conventional selection methods lead to confidence intervals that are too narrow and p values that are too small. Furthermore, theory and simulation evidence have found no selection method to be uniformly superior to adjusting for all well-measured confounders. Nonetheless, control of all measured confounders can lead to problems for conventional model-fitting methods. When these problems occur, one can apply modern techniques such as shrinkage estimation, exposure modeling, or hybrids that combine outcome and exposure modeling. No selection or special software is needed for most of these techniques. It thus appears that statistical confounder selection may be an unnecessary complication in most regression analyses of effects.

Bayesian methods; collapsibility; confounding; epidemiologic methods; regression; shrinkage; validity; variable selection

Epidemiologists often have available a set of many potential confounders for a targeted exposure-disease relation and may attempt to select or delete variables from this set using statistical methods. Hoffmann et al. (1) give a method based on a collapsibility test, and there are several other collapsibility tests that can be applied in the same way (2–6). For example, earlier authors (3, 4) gave much simpler collapsibility tests based on differencing the adjusted and unadjusted variances, which they showed can be applied to odds ratios and to rate ratios (4) and which extend to generalized linear models, including logistic and log-linear rate models (5, 6).

These methods may be useful when the primary research question is whether control of a set of covariates makes an “important” difference in settings like the one at hand. Yet, more often this question is a preliminary one, asking which variables should be used for confounding control in the current analysis. Are such variable-selection methods still needed for this task? In the past millennium, the answer was “yes,” given the severe limitations that existed on the number of variables that could be controlled in one model (7). Nonetheless, their application was always theoretically defective in failing to account for selection effects on subsequent tests and confidence intervals for the exposure effect. Fixing this defect is not simple.

Fortunately, statistical confounder selection may no longer be necessary in typical applications. No selection strategy has proven uniformly better than adjusting for all well-measured confounders after preliminary screening to remove doubtful or inappropriate covariates. Statistical problems from controlling many covariates can be addressed by using modern adjustment techniques, such as shrinkage estimation and exposure modeling. If one wishes to delete covariates to simplify analysis and presentation, the safest approach may be to allow deletion only if it has negligible impact on the final confidence interval for the exposure effect.
CONFOUNDER SELECTION

Background screening of candidates

Background knowledge usually identifies covariates that are best controlled no matter what (typically age and sex) or matching factors in case-control studies that predict exposure strongly (8). Nonetheless, many other covariates can be excluded from control on the basis of background knowledge, for example, because they are affected by exposure or disease (8–10). Causal diagrams provide useful rules for such background screening, including rules for identifying covariate sets that are minimally sufficient for confounder control, assuming the diagram is correct (9, 10).

Some covariates that pass this background screening might be excluded because they are judged to have too little variation, too many missing data, or too much measurement error to be worth including. Unfortunately, there are as yet no well-founded guidelines for making these judgments. Nonetheless, such preliminary concerns are best raised early enough to influence data collection, for collecting too many data may tax quality control and subject cooperation, thus degrading the accuracy and completeness of data while increasing drop-out rates.

The structure of confounder selection

If background screening identifies Z as a confounder, the only theoretical rationale for deleting Z is that the resulting bias is negligible or is more than offset by a reduction in variance. Most confounder-selection methods do not address the bias-variance trade, however, and instead focus only on the size of the bias. To describe these methods, suppose we have a logistic risk model or log-linear rate model (such as a proportional-hazards model) in which θ is the coefficient of a covariate Z, and βa is the corresponding “Z-adjusted” exposure coefficient. Also, let βu be the “unadjusted” exposure coefficient in the model without Z. The θ, βu, and βa are parameters, representing what would be obtained in a sample so large that random error was negligible.

If βa ≠ βu (equivalently, if βa − βu ≠ 0), the exposure coefficient is noncollapsible over Z, and βa − βu measures the degree of noncollapsibility. Confounding by Z is often measured by the change in exposure-disease association upon adjustment for Z, βa − βu, or a function of it such as the ratio of adjusted and unadjusted ratios exp(βu)/exp(βa) = exp(βa − βu). These measures can be useful approximations in some settings, but can be misleading, as will be discussed below. The remaining discussion will focus on βa − βu but applies to other measures as well. It also applies if Z is a list (vector) of covariates (1, 5, 6).

Significance testing

Although common, selection based on testing the Z coefficient θ is logically flawed, in part because θ is just one factor that determines the degree of confounding by Z (7–13). The latter problem is addressed by testing collapsibility instead (1–6).

Significance testing of collapsibility selects Z if the null hypothesis βa = βu (collapsibility) is rejected (2), that is, if the observed change βa − βu is “significant.” Examples and simulations have found, however, that use of p < 0.05 for “significant coefficient” or “significant change” (α = 0.05) too often leads to deletion of important confounders (false negative decisions) (14–17). Use of much higher α levels for the confounder test (e.g., α = 0.20, selecting if p < 0.20) mitigates this problem (15–17).

A more general objection, however, is that significance testing treats false negative error (deleting a confounder) as secondary to false positive error (selecting a harmless nonconfounder). It has been argued that this ranking of error is backward: Deleting a confounder introduces bias and thus can be justified only if one can infer that the bias is tolerably small or worth the precision gain, whereas retaining a harmless nonconfounder only reduces precision (which is the price paid for protection against confounding by Z) (7, 11–14).

The objection to significance testing may be seen more clearly by using confidence intervals. Deleting Z when the test of βa = βu is nonsignificant is the same as deleting Z if the confidence interval for βa − βu includes zero. In small studies, the interval will be wider and thus more likely to include zero; hence, significance testing is more likely to delete covariates in small than large studies, even if the covariates are important confounders (12, 13, 16). The same interval may also include large differences, in which case nonsignificance of the observed difference βa − βu provides no assurance that the true difference βa − βu is small.

Equivalence testing

To obtain some assurance that Z can be safely deleted, some authors consider equivalence testing, in which Z is deleted only if the confidence interval for βa − βu includes no large value (1, 5, 17). For example, with use of −0.1 to 0.1 as the range of tolerable bias from deleting Z (corresponding roughly to 10 percent bias on the odds-ratio or rate-ratio scale), Z would be deleted only if the interval for βa − βu fell in this range (1, 5, 17). With use of a 95 percent or even an 80 percent confidence level and a 10 percent bias tolerance, equivalence testing may delete few covariates or none in smaller studies; this is because it cannot delete a covariate unless the interval for βa − βu is narrower than the range of tolerable bias.

More often, Z is deleted if the change in the point estimate, βa − βu, falls in the designated tolerable range (11–14). This approach corresponds to using a 0 percent confidence interval for equivalence testing. Some authors recommend instead deleting Z only if the difference in the confidence intervals for βa and βu appears negligible (8, p. 258; 14). This refinement limits the impact of deletions on inferences about the exposure effect.

Parallels and differences in the methods

Significance and equivalence testing are based on the same computation, that is, an estimate βa − βu of βa − βu, and its variance estimate. For stratified analyses or for small βa − βu (which are the focus of equivalence testing), an approximate variance for βa − βu is just the variance of βa
minus the variance of $\hat{\beta}_b$ estimated from models with and without $Z$, that is, the adjusted minus the unadjusted variance (3–5). Clogg et al. (6) give a more general variance formula, while Hoffmann et al. (1) give another formula that represents $\beta_a$ and $\beta_b$ within a single model. All of these formulas extend to vector $Z$. Nonetheless, at this time there is no evidence regarding the relative performance of the different collapsibility tests that result.

As mentioned above, significance testing (whether of $\theta = 0$ or of $\beta_a = \beta_b$) is more likely to delete variables in smaller samples, where many variables will not be clearly confounding; in contrast, equivalence testing is more likely to delete variables in larger samples, where some variables may clearly emerge as unimportant. Because variances are larger, unnecessary control is of greater concern in small samples. These considerations arguably favor significance testing over equivalence testing. Many other criteria are possible. Nonetheless, limited simulations thus far have not found a confounder-selection method that always outperforms others when the criteria levels and bias tolerance are set to limit serious false negative error (16, 17).

**Addressing distortions produced by selection**

In practice, variable selection is dogged by the fact that the final variance estimates tend to be downwardly biased if they do not account for the selection, while the point estimates may suffer related distortions (18–24). Various procedures have been proposed to address these problems (21–24). A simple fix involves resampling (“bootstrapping”) the data with replacement and then repeating the entire analysis procedure, including the variable selection. Confidence limits and $p$ values can be computed from the resulting simulation output (25, 26). Unfortunately, simply using percentiles or the variance of the simulated exposure-disease associations (which is common practice) can give inaccurate confidence limits and $p$ values (25, 26), especially when selection is part of the process (20). All told, it is difficult to construct valid statistics that account for selecting variables or model forms on the basis of the data (23, 27).

**Problems with the noncollapsibility criterion**

A problem for any analysis is that $Z$ may not be a confounder even if the exposure-disease association is noncollapsible ($\beta_a \neq \beta_b$) and $Z$ precedes exposure and disease. That is, $Z$ may look like a confounder both statistically and temporally, but its control may only worsen bias. One way this can happen is if the relations of $Z$ to the exposure and disease are themselves confounded (9, 10). For example, under figure 1 with $Z$ binary, upon control of $Z$ we would find $\beta_a \neq \beta_u$; nonetheless, it is the adjusted coefficient $\hat{\beta}_a$ that is confounded (albeit by $A$ and $B$ rather than by $Z$).

For logistic and proportional-hazards models, noncollapsibility can occur even if there is no confounding of *any* relation (28–33). For example, if one uses maximum-likelihood estimation, the observed change $\hat{\beta}_a - \hat{\beta}_b$ may represent only increased sparse-data bias in the adjusted estimate $\hat{\beta}_a$ (30–33). Thus, collapsibility testing can lead to unnecessary or even harmful control based on statistical artifacts. Furthermore, with a common outcome, noncollapsibility may reflect only the inequality of marginal and stratum-specific effects, rather than confounding (28, 29).

**CONTROLLING ALL MEASURED CONFOUNDERS**

To minimize the risk of confounding and to avoid the problems and complexities of variable selection, after preliminary screening we might fit a model regressing incidence or prevalence on exposure and *all* the measured potential confounders. As mentioned above, however, conventional fitting methods such as maximum likelihood (whether unconditional or conditional) may then suffer from estimate inflation (sparse-data bias) or may fail to converge. These problems are likely if there are too few cases or noncases per model coefficient, or if the cases or noncases are too concentrated at single covariate values (30–34). Nonetheless, inflation may occur even if there is only one confounder (32, 33).

Real examples have exhibited large inflation of odds ratios because of sparse data, often unrecognized in the original reports (32–36). To control inflation and introduce background information into the model, one can use shrinkage methods, also known as random-coefficient, ridge, Stein, empirical-Bayes, semi-Bayes, hierarchical, multilevel, and penalized estimation (8, chap. 21; 24, 31–35, 37–54). These methods pull (“shrink”) coefficient estimates toward zero or more generally toward values expected from background information (Bayesian analysis) or estimated from the data (empirical Bayes). Given a covariate $Z$ with coefficient $\theta$, the pull is proportional to the estimated variance of $\theta$; thus, unstable estimates are shrunk more than stable ones.

Variable selection can be viewed as crude shrinkage in which $\hat{\theta}$ is either pulled all the way to zero (deleted) or else not shrunk at all (19). Shrinkage methods allow more flexibility by letting a variable be partially deleted to an extent determined by the instability of $\hat{\theta}$, although some varieties (e.g., the Lasso (24)) may completely delete certain
variables. Related approaches shrink the adjusted exposure-coefficient estimate $\beta_2$ toward the unadjusted estimate $\beta_1$ in order to minimize the estimated mean-squared error (17, 55).

The theoretical advantages of shrinkage over conventional methods have been confirmed in many real examples and simulation studies (24, 32–35, 38–54). Of great practical importance, however, is that shrinkage can be done with ordinary software for logistic, Poisson, or Cox models via the method of prior data (35, 41, 48, 53, 56–58).

**Shrinkage via prior data**

Shrinkage estimation with prior data first appeared in the 18th century (41). To describe the original method of Laplace, suppose we wish to estimate the probability of disease $\pi$ from a random sample with $A$ cases and $B$ noncases out of $N = A + B$ total. The method adds one case and one noncase to the data, thus shifting from the maximum likelihood estimate $A/N$ of $\pi$ to the shrinkage estimate $(A + 1)/(N + 2)$, which is closer to $\frac{1}{2}$. Similarly, the estimate of $\theta = \logit(\pi) = \ln[\pi/(1 - \pi)]$ is shifted from the maximum likelihood estimate $\ln(A/B)$ to $\ln[(A + 1)/(B + 1)]$, which is closer to $\logit(\frac{1}{2}) = 0$. Thus, Laplace’s method shrinks the estimate of $\theta$ toward zero by adding two “prior observations” to the data. The degree of shrinkage depends on the sizes of $A$ and $B$, becoming important only if $A$ or $B$ is small.

Laplace’s method can also be applied when $\theta$ is the coefficient of $Z$ in a logistic risk or a Poisson (log-linear) rate model (48): We add two “prior-data” records to the data, one with $Z = 1$ and the other with $Z = 0$. Each record has one case and 100,000 persons or person-years, with all other variables set to zero. These two records are a stratum of “prior data” with an odds ratio or rate ratio of 1, giving an estimate of $\ln(1) = 0$ for the log ratio $\theta$. This prior stratum pulls the estimate of $\theta$ toward zero. To avoid confounding of the prior data with the actual data, we also add an indicator variable Prior$_{Z}$ to the regression, with Prior$_{Z} = 1$ for the two added records and $= 0$ for all other records.

Using just one case per prior record corresponds to 95 percent prior certainty that $\exp(\theta)$ is between $1/39$ and 39 and has little impact on most coefficients. If we wish to narrow the probable range for $\exp(\theta)$ and thus induce more shrinkage, we add more prior cases. For example, the use of four cases per record corresponds roughly to 95 percent prior certainty that $\exp(\theta)$ is between $1/4$ and 4, whereas the use of 16 cases per record corresponds to 95 percent prior certainty that $\exp(\theta)$ is between $1/2$ and 2 (56).

There are many other extensions of the prior-data method. To pull the estimated $\theta$ toward a nonzero number $\theta_0$, we set the denominator in the second record to $\exp(\theta_0) \times 100,000$ (56). To pull more in one direction than another, we skew the prior data (53, 58). We add two prior records and one prior indicator for each coefficient we wish to shrink; the number of cases added and the ratio of denominators may differ among coefficients (48). To shrink coefficients in conditional-logistic or proportional-hazards regression, however, we add counts of discordant matched pairs instead of counts of individuals, and we need no prior indicator (48, 57).

For $\exp(\theta)$ and the prior data to make sense, 0 and 1 must be meaningfully different values for $Z$, which may require recoding of $Z$. For example, if $Z$ is diastolic blood pressure in millimeters of mercury, we may subtract 80mm to make $Z = 0$ correspond to 80mm and then divide by 10 to make $Z = 1$ correspond to 90mm; $\exp(\theta)$ would now be the odds ratio or rate ratio comparing 90mm ($Z = 1$) with 80mm ($Z = 0$).

**Shrinkage for control of collinear confounders**

Unlike conventional methods, shrinkage can be applied when the covariates are collinear and conventional methods fail completely. In particular, shrinkage allows control for potential confounding between primary recorded variables and derived quantities, such as confounding between food and nutrient effects.

As an example, table 1 shows the results of applying various methods to a study of 140 breast cancer cases and 222 matched controls (34, 59). Nutrient intakes were derived from a food questionnaire using composition tables; hence, the 35 derived nutrient measurements are collinear with the 87 food items. Keeping all 35 nutrients in the conditional logistic model identifies omega-3 fatty acids and phytoestrogens as negatively associated with breast cancer and beta-carotene as positively associated. Stepwise (backward) deletion using $\alpha$-to-remove of 0.10 only slightly changes these results.

Although such analyses go unquestioned in the literature, they are not credible insofar as they assume that all the food coefficients are zero. That is, they assume that each food $Z$ has no residual effect (no effect beyond that due to the measured nutrients). These assumptions are necessitated by the fact that conventional methods cannot fit a model with collinear covariates and so cannot include all the foods as confounders. This limitation leaves them vulnerable to large net confounding: Even if none of the omitted variables has a large effect by itself, their net effect taken together could be considerable. This possibility is a greater validity threat to the stepwise analysis, because it also assumes that most nutrient coefficients are zero.

Shrinkage methods allow one to address residual confounding by entering each food in the model, along with prior data that limit the size of the residual food effects. In the example, entering 16 case-control pairs with $Z = 1$ for the case and $Z = 0$ for the control, as well as 16 pairs with $Z = 0$ for the case and $Z = 1$ for the control, is equivalent to assuming a normal prior for the $Z$ coefficient that places 95 percent probability on the odds ratio $\exp(\theta)$ being between $1/2$ and 2 (48). When all 87 foods are entered in the regression along with these prior data, the nutrient results look far less certain than they do in the conventional analysis; refer to the final column of table 1.

The omission of foods from the conventional analyses corresponds to the extreme of betting 100 percent that every food coefficient is zero. The same result would be obtained by entering each food with an infinite number of matched pairs as prior data. The omission of an additional 20 nutrients by the stepwise analysis corresponds to extending this unsupportable certainty to most of the nutrients. The unwarranted certainty that residual food effects are absent translates into overly narrow confidence intervals for the nutrient effects.
Such excessive certainty in conventional results may help to explain refutations of epidemiologic findings by randomized trials (60). Extensions of shrinkage methods to bias modeling allow incorporation of further sources of uncertainty, such as selection bias, misclassification, and unmeasured confounders, leading to greater caution in inferences about highly “statistically significant” or precise-looking associations (61).

**EXPOSURE MODELING**

Many clinical studies have good information for treatment prediction, which allows improved confounder control via treatment (exposure) modeling. Methods that use exposure modeling include adjustment by exposure regression (11, 62, 63), exposure probability (propensity scoring) (64, 65), inverse-probability weighting (65, 66), and hybrid methods that use both exposure and outcome models for adjustment (67–69).

Although exposure modeling is a burgeoning topic, present methods have difficulties and need more development, especially for case-control analysis. For example, exposure-effect estimates from exposure modeling can suffer variance inflation if the fitted model predicts exposure very well, even if no confounding is present. Again, screening out nonconfounders on subject-matter grounds is essential, and statistical variable selection among what remains is problematic. Collapsibility tests based on exposure models do not yet exist, while methods that test only the covariate-exposure association can easily delete important confounders, yet worsen variance inflation by selecting nonconfounding but strong exposure predictors (70). Relevant avenues for investigation include development of valid methods for variable selection and shrinkage in exposure and hybrid modeling.

Exposure modeling does have a sample-size advantage in cohort studies when the outcome is rare but exposure is common, for then exposure risk can be more stably estimated than outcome risk (71). This advantage does not lead to greater efficiency, however, because the final precision remains limited by the small numbers of cases. The sample-size advantage can shift to outcome modeling when the outcome is more common in the study than the exposure, as in many case-control studies. Furthermore, exposure modeling is prone to several statistical artifacts in case-control studies (72).

**CONCLUSION**

Examination of the results of different analytical approaches is helpful, especially when one understands the strengths and weaknesses of each method. Confounder selection methods have numerous weaknesses that should encourage direct examination of results from control of all measured confounders when that is feasible, even if only for reference. When full control leads to problems for conventional regression programs, one can apply modern techniques such as shrinkage, exposure modeling, or hybrid methods as outlined above, rather than rely on conventional variable-selection algorithms that can easily delete or ignore important confounders. Versions of these techniques can be performed with common software packages, and they can provide more accurate estimates and uncertainty assessments than can conventional methods. Although no statistical method is a panacea for epidemiologic problems, these

---

**TABLE 1. Estimates of odds ratios from conditional logistic regressions of breast cancer on nutrients (food constituents)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Conditional maximum likelihood on nutrients</th>
<th>Stepwise deletion of nutrients</th>
<th>Shrinking food coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds ratio</td>
<td>95% confidence limits</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>Omega-3 fatty acids (g/day)</td>
<td>0.71</td>
<td>0.46, 1.1</td>
<td>0.77</td>
</tr>
<tr>
<td>Beta-carotene (mg/day)</td>
<td>1.2</td>
<td>1.01, 1.3</td>
<td>1.1</td>
</tr>
<tr>
<td>Phytoestrogens (mg/day)</td>
<td>0.73</td>
<td>0.58, 0.93</td>
<td>0.80</td>
</tr>
<tr>
<td>Alcohol (3 ounces/day)</td>
<td>0.89</td>
<td>0.63, 1.3</td>
<td>0.94</td>
</tr>
<tr>
<td>Carbohydrate (100 g/day)</td>
<td>0.97</td>
<td>0.79, 1.2</td>
<td>1</td>
</tr>
</tbody>
</table>

* Five potential confounders are forced into each model plus 87 foods (dietary items) in the shrinkage model.
† Includes all 35 nutrients but no food item. Equivalent to adding an infinite number of pairs for each food item.
‡ \( \alpha \)-to-remove = 0.10; 15 nutrients retained and 20 deleted. Equivalent to adding an infinite number of pairs for each food and each omitted nutrient.
§ Includes all 35 nutrients (food constituents) and all 87 foods (diet items); shrinks food coefficients toward zero by adding 32 discordant pairs per food (16 with a case-control difference of 1 unit for the food, 16 with a difference of \( -1 \) unit, and all other covariates zero), adding a total of \( 32 \times 87 = 2,784 \) pairs.
¶ One ounce \( = 29.6 \) ml.
methods deserve to become part of the regular curriculum in epidemiologic analysis.

ACKNOWLEDGMENTS

The author thanks Babette Brumback, Katherine Hoggatt, and Charles Poole for helpful comments. Conflict of interest: none declared.

REFERENCES