Statistical Foundations for Model-Based Adjustments

Sander Greenland[1] and Neil Pearce[2,3]

[1]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, California 90095-1772; email: lesdomes@ucla.edu

[2]Departments of Medical Statistics and Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom; email: Neil.pearce@lshtm.ac.uk

[3]Centre for Public Health Research, Massey University, Wellington 6140, New Zealand

Annu. Rev. Public Health 2015. 36:89–108

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

This article's doi: 10.1146/annurev-publhealth-031914-122559

Keywords

causal inference, confounding, modeling, variable selection

Abstract

Most epidemiology textbooks that discuss models are vague on details of model selection. This lack of detail may be understandable since selection should be strongly influenced by features of the particular study, including contextual (prior) information about covariates that may confound, modify, or mediate the effect under study. It is thus important that authors document their modeling goals and strategies and understand the contextual interpretation of model parameters and model selection criteria. To illustrate this point, we review several established strategies for selecting model covariates, describe their shortcomings, and point to refinements, assuming that the main goal is to derive the most accurate effect estimates obtainable from the data and available resources. This goal shifts the focus to prediction of exposure or potential outcomes (or both) to adjust for confounding; it thus differs from the goal of ordinary statistical modeling, which is to passively predict outcomes. Nonetheless, methods and software for passive prediction can be used for causal inference as well, provided that the target parameters are shifted appropriately.

INTRODUCTION

In our experience, few topics cause as much consternation, both in students and in researchers, as modeling strategies do. It is straightforward to do a course on the basic principles of confounder control and assessment of interaction and mediation and to learn how to run regression models. It can be far more daunting to be confronted with a new data set that includes many potentially important covariates and have to decide what to do with it.

Most epidemiologic textbooks are vague on the practicalities of model selection, and understandably so. Arguably, model selection should be strongly influenced by factors that are specific to the particular study, including possibly controversial prior information about which variables are important potential confounders, modifiers, or mediators. It can thus be counterproductive to impose rigid modeling rules or recipes. Nonetheless, lack of guidelines can leave in place bad practices, such as choosing models based on naïve significance testing. General recommendations may thus be useful even if exceptions are common. Similarly, critical assessment of analyses requires knowledge of which strategy or guidelines were used to select analysis models.

The present article updates older commentaries on modeling guidelines (46, 47, 116, 139). Our motivation stems from noting that few subject-matter studies employ newer modeling methods (despite many sophisticated papers on these methods in leading epidemiology and statistics journals) and that simple but problematic methods remain common defaults in basic teaching, commercial software, and the clinical literature. Although simple methods may often suffice, it is important to understand their limitations and to recognize when better alternatives are required.

Scope, Aims, and Assumptions

We focus entirely on methods for observational studies of causal effects, there being some excellent texts on purely predictive modeling (79, 80, 94, 122, 132). We begin by outlining the precepts we take as fundamental to sound use of statistical methods in epidemiology, emphasizing the importance of understanding the contextual interpretation of model parameters, selection criteria, and estimates. We then critically review some established modeling approaches based either on passive predictive models or on changes in the estimate (CIE) of exposure effect and some variations of these approaches available to researchers with more technical resources. Our coverage is not intended to be a comprehensive review for highly skilled practitioners; rather, we target teachers, students, and working epidemiologists who want an accurate data analysis, but who lack resources such as R programming skills or a bona fide expert both in their field and in statistical modeling committed to their project. Elsewhere (59, 134) we discuss how simple approaches can be recast and upgraded, with little effort and without special software, to minimize harmful practices and follow more sound methodologic principles.

Throughout, we assume that we are applying a conventional risk or rate regression model (e.g., logistic, Cox, or Poisson regression) to estimate the effects of an exposure variable on the distribution of an outcome variable, while controlling for other variables ("covariates"), or that we are applying an exposure model for adjustment (as in propensity score methods). The covariates may be forced variables, which we always want to control (typically for age and sex), or unforced variables for which the decision to control may be data based. Our focus is on strategies for decisions about unforced variables.

The strategies we discuss apply to exposures and covariates of any form. We assume that data checking, description, and summarization have been done for quality control (73, 150); thus, we do not address problems arising from preliminary data examinations (e.g., from influence

on subsequent analyses through collapsing away small categories) (18, 19). We advise using univariate distributions and background (contextual) information to select categories or an appropriately flexible form (e.g., splines or fractional polynomials) for detailed modeling of quantitative variables (5, 61, 67), but we must leave many difficult issues about model specification and diagnostics to more detailed discussions (3, 60, 61, 67, 79, 80, 94, 122, 132, 140). Finally, we do not consider special models and problems that arise in the contexts of ecologic analysis (55, 151), time-varying exposures (83, 140), mediation analysis (83, 115, 117, 138, 140, 142), measurement-error adjustment (17), or bias analysis (64, 68, 69, 77, 78, 92, 93, 137), although our general comments apply to all modeling applications.

OUR FUNDAMENTAL STATISTICAL PRECEPTS

Why Contextual Interpretation and Sensibility Should Dominate Statistical Theory

We agree with other authors that scientific theories, whether concerning substance or concerning methodology, are not uniformly reliable and are often filled with fictions and falsehoods (39, 113, 159). This caution applies to statistical theories of inference; in particular, strict adherence to a statistical philosophy or mathematical theory can harm scientific inference unless such adherence is moderated by contextual sensibility (13, 43, 88, 107, 111, 128, 135). In messy observational sciences such as epidemiology, this harm is often traceable to reification (157) of probability distributions and their properties (that is, treating hypothetical distributions as if real). Damage to scientific inference is especially acute when reification leads to accepting conclusions from methods whose properties are defined only within idealized sequences or models. The most common examples involve treating nominal error rates of tests as if they were real (e.g., taking a 0.05 α level as guaranteeing no more than 5% of rejections are incorrect when the tested hypothesis is correct); these error rates are in fact derived from distributions representing hypothetical infinite sequences of valid studies (which do not exist among real observations in health and medical sciences) and usually assume that there is no data-based model selection.

Observational epidemiologic studies have unique source populations, protocols, and unanticipated problems that exemplify the tenuous nature of frequency models. Consider the Nurses Health Study I (NHS I) (23): There is no similar cohort of nurses observed in such detail during the final quarter of the twentieth century, a period when low-fat, high-carbohydrate diets remained a common recommendation for weight-control and health, when statins were not yet widespread, and so on. The study was thus conducted on a population inaccessible to future studies. Available comparison cohorts do not and never will resemble the NHS-I cohort with respect to distributions of all important health predictors and measurement errors, so there is no guarantee that fully effective adjustments can be made for differences. Consequently, the number of cohorts credibly resembling NHS I after adjustment will be so few that analyses assuming replication with only random differences will be purely hypothetical.

In sum, standard statistical theories model only random uncertainty, i.e., uncertainty about what would emerge from attempts to replicate the study methods down to the last known detail predictive of outcome, successfully controlling for all other uncertainty sources (biases). Such detailed valid replication is impossible outside of a perfect parallel randomized trial. Health, medical, and social sciences rarely have such replication, and their total uncertainty often greatly exceeds uncertainty from random variation [as illustrated in bias analyses (68, 64, 69, 77, 78, 92, 93, 137)].

Case Studies in Reification: Statistical Unbiasedness and Consistency

Properties defined only over long runs must appeal to hypothetical sequences under models that are grossly oversimplified relative to the forces determining background risks, participation, and measurement errors (57). Although statistical significance and claims of objectivity are the usual targets of this criticism (9, 48, 58), concepts of unbiased and consistent estimation are also vulnerable. These concepts refer to estimators whose mean and large-sample limit equal the targeted parameter under an assumed repeated-sampling model.

A major problem is that the models used to show unbiasedness or consistency are known to be false: Uncontrolled epidemiological biases are always present and render inoperative theoretical guarantees of unbiasedness and consistency for the actual target parameters (57). Thus, no available estimator can be shown to be unbiased or consistent under realistic epidemiologic conditions. Even when only random error is present, however, unbiasedness is neither necessary nor sufficient for contextually sensible estimators (16, 33, 52, 97). We thus reject unbiasedness as an absolute requirement for point estimators of target parameters.

The hypothetical nature of distributions does not make statistical theory or methods useless, but it does greatly reduce the decisiveness and accuracy of the methods to below that indicated by *p*-values or confidence intervals. Properties under hypothetical distributions may serve as useful guides for determining when the distributions appear contextually reasonable. Nonetheless, any restriction of judgments by theoretical rules risks enormous error that might have been obvious from a contextual viewpoint [e.g., failure to adjust a covariate simply because it is not significant in a risk model despite its known role in etiology (71)]. Thus, statistical theories must be used with the understanding that they are theories of unattainable ideals. If a researcher has a sound basis for the assumptions of the theory and a sound mapping of these assumptions into the research context, deductions within the theory (its theorems and methods) can serve as analysis directives (87); for example, if only random error is present, statistical consistency is a compelling requirement for an estimator. But given uncertainties about the assumptions and the mapping, the use of a statistical theory as if it were certainly correct cannot be justified, and another rationale for its use should be sought.

Uses of Statistical Theory and Modeling

One rationale for statistical theories is that they can be used to conduct valuable (albeit simplistic) thought experiments within their narrow confines (77). We can analyze data using different statistical theories and methods and compare the inferences so obtained. Similarly, we can evaluate and compare methods within theories, and we can compare these evaluations across theories. For example, we can compare modeling strategies within and across frequentist theories (whether Neyman–Pearson or Fisher–Cox) and Bayesian theories (whether likelihood-based or personal betting), regardless of which theory originated them. We can also compare their performance in passive prediction versus causal (interventional) prediction. Doing so, we educate our intuitions regarding when criteria and methods will be misleading or useful, on the basis of how well the scientific problem can be mapped into the theory.

As an illustration, the unbiasedness criterion can be criticized even within the confines of classical frequentist criteria for comparing inferential methods. The criterion assumes that systematic error in a coefficient estimator is always far more costly than random (mean-zero) error. Consequently, classical unbiased estimators often perform much worse than do biased estimators according to practically relevant valuations, such as when error costs are proportional to total error or when out-of-sample prediction is at stake (which is often evaluated in terms of mean squared error or mean absolute error) (16, 25, 33, 35, 36, 52).

More generally, different strategies can be justified by different assumptions, and no strategy can be proven best or optimal or even sound when its assumptions are not assuredly true. Furthermore, criteria for optimality vary with goals and valuations. Some strategies may be preferred over others based on theory; however, these preferences arise only in special cases in which the discrepancy between statistical and contextual theory is judged minor. Thus in practical terms there are no best or optimal strategies, and the reader should be suspicious of any method or model promoted as optimal or correct. Every model and every strategy has limitations and defects that depend on the context in which it is applied; thus, given enough analysis time, every model would only contribute one result among many in a sensitivity analysis over models.

Both frequentist and Bayesian methods are subject to these cautions, because both frequency (error) calibration and Bayesian (betting) coherency are unattainable ideals in health and social research. In practice, the data models used by both methodologies fall far short of capturing all sources of error and uncertainty. Thus frequentist and Bayesian methods are simply alternative ways of looking at models and data, and are more complementary than competitive (8, 34, 43, 58, 77).

In summary, although mathematical investigations and simulations are invaluable for formulating guidelines for judgment, they rely on simplifying assumptions that are almost never fully satisfied and thus can be misleading if their results are taken with the deductive certainty obtained under their assumptions (66). On the other hand, counterexamples can show how guidelines can fail but must be critically evaluated with respect to their contextual realism and importance (66). Beyond theoretical investigation, it is equally important to assess and compare modeling strategies in real case studies involving various conditions to see when in practice they yield contextually sensible results.

SOME PRACTICAL PRELIMINARIES FOR SOUND MODELING

Unfortunately, analysis resources are usually severely limited, and many principles will be compromised to deliver the results on schedule. As a consequence, model checking may be absent or limited in scope, which makes essential a good grasp of the contextual meaning of a model and its fitting method in spotting model defects. In particular, a good modeler will recognize each assumption and valuation enforced by the model and judge these according to the degree they are contextually supportable (corroborated by available information), uncertain (plausible but not well corroborated), or implausible (contradicted by available information). The modeler may decide to use uncertain or even partially implausible models, which may be defensible if one is aware of the misfits between the model and the contextual information. The most thorough analyses will, however, probe uncertain assumptions with sensitivity analysis, discussing the contextual status of each model (e.g., supportable, uncertain, or implausible).

Any apparent conflict between formal methods and informal assessments may stem from faults in our expectations, faults in the formal methods, or both. No one doubts that prior expectations can be quite faulty, distorted as they are by biases in the methodology, interpretation, and reporting of previous studies (63, 90–93, 112). But formal statistical inferences can be quite faulty as well, insofar as they fail to account for important background information or uncertainty (error) sources such as cognitive biases built into the methodology (90–93). Conflicts will require resolution because they indicate an error in one or both of the statistical and contextual theories.

Crucial Data Preprocessing: Meaningful Centering and Scaling

There are two data-processing steps that are important to contextually sensible modeling but are usually neglected: Quantitative variables should be recentered to ensure that zero is a meaningful

reference value present in the data, and they should be rescaled so that they are measured in general scientific units that represent meaningful differences spanning a range present in the data (61). Consider diastolic blood pressure (DBP) recorded as millimeters above zero (no pressure). This is a clinically unintelligible scale: A 1-mm change in pressure is below measurement noise and clinical significance, and the reference (zero) point corresponds to death. Any effect estimate (such as a regression coefficient) presented in these terms would thus be difficult to interpret because no one has a good intuition for the effect of a 1-mm change. To remove these difficulties, DBP could be recentered so that 0 represents 80 mm (often taken as clinically normal) and then rescaled to centimeters instead of millimeters so that 100 mm would become (100–80)/10 = 2 cm above reference level; now the coefficient would refer to a 1-cm (10-mm) increase in DBP from a recognizable clinical reference point. Similarly, in a study of older adults, age could be recentered so that 0 represents age 60, and rescaled to decades instead of years, so that 80 years would become (80–60)/10 = 2 decades past 60.

Unfortunately, the statistical literature is replete with bad scaling recommendations and practices, using arbitrary study-specific quantities such as sample standard deviations or interquartile ranges, despite lengthy critiques of these practices (70, 75). To illustrate, suppose the rate ratio relating packs/day smoking to mortality was 3 in every study. Using standard deviation (SD or standardized) units, each study would use a different (and strange) unit to measure smoking: A study in a population with high smoking prevalence and a smoking SD of 0.67 packs/day would report an exponentiated coefficient of $3^{0.67} = 2.1$, whereas a study in a low-prevalence population and a smoking SD of 0.25 would report an exponentiated coefficient of $3^{0.25} = 1.3$; thus the effect in the latter study would appear much weaker, even though the individual effect of smoking is identical in both populations. Thus, converting to SD units does exactly the opposite of standardization in ordinary English terms because it removes each study from a contextually understood standard (packs/day) and instead produces different units in each study. SD units will often not even apply to the source population of the study itself; for example, matching alters the sample SD of any variable associated with matching factors.

Special Cautions Regarding Product Terms

Recentering and rescaling are especially important for coefficient interpretability when examining effects of exposure combinations (interactions) (61, 100) and also when using prior distributions or penalty functions (59, 134). To illustrate, suppose we observed a cohort in which smoking conferred a mortality rate ratio (RR) of 3 when expressed in packs/day, but smoking was fitted as cigarettes/day; then, with 20 cigarettes/pack, the exponentiated smoking coefficient (output RR) in cigarettes/day would be $3^{1/20} = 1.056$. Suppose also that systolic blood pressure (SBP) had a mortality RR of 4 for a 40-mm range but is fitted in millimeters; then the exponentiated SBP coefficient would be $4^{1/40} = 1.035$. In a multiplicative model with these conditional effects, the combined effect would be $3^{1/20}4^{1/40} = 1.094$ and thus appear unimportant, despite the huge RR of 3(4) = 12 when comparing pack-a-day smokers to nonsmokers with 40-mm lower SBP.

In nearly all contexts, models containing a product term should also contain the factors in the product as main effects (the hierarchy principle) (61, 100). Without a component main effect, the product-term coefficient will depend on the center (reference or zero point) chosen for the other factor in the product, which complicates correct interpretation.

Preliminary Screening and Posterior Diagnostics Based on Background Information

Perhaps the most well-known and accepted way that contextual information enters a causal analysis is through preliminary identification of confounders based on established causal relations (44, 121). For example, estimation of net (total) effects typically requires exclusion of intermediates and their effects (44, 71, 72, 108) and any other variable influenced by the exposure or outcome (24, 72, 108, 121, 141), as well as other variables whose control may increase bias without reducing total error (24, 44, 72, 103, 108, 109, 121, 141). We assume that these variables have been identified and eliminated, leaving us with potential adjustment covariates (potential confounders). These include covariates considered essential to control (e.g., age, sex), which must be forced into models along with the study exposures. They also include covariates we are confident would reduce bias if controlled properly if only our study size were unlimited, as well as covariates of uncertain adjustment value.

In this planning stage of analysis, it is valuable to record what one expects to see for each coefficient (which is another reason why meaningful scaling is important). After the data are analyzed, one can compare these expectations to the estimates. Seemingly large discrepancies may indicate error in one's expectation, error in the analysis model, error in the data, or some combination. This contrast of residual expectations (those not used in construction of the analysis model) against model-fitted expectations (those deduced from the data model and the data) is thus an error diagnostic. The contrast can play this role even without formal testing, but formal tests can be constructed if our expectations can be assigned some degree of certainty in the form of a prior distribution or external data (12, 58).

As an example, take age in relation to typical carcinomas. From vital statistics, one expects cohort curves for these outcomes to be very steeply positive. If one then obtains a cohort coefficient of age (in decades) that is near or below zero, this apparent anomaly requires an explanation. Of course, in a very small study this difference might be within conventional random-error bounds; but if random error is not a plausible explanation, a causal explanation for the anomaly will be needed. Such anomalies may warn of problems with the study or the data model or may reflect errors in our background information.

Expectations can also be mistaken if they fail to account for features of study design and execution. For example, in an age-matched cancer case-control study, the relation of age to cancer will become sawtoothed, with jumps at the ends of each matching category, rendering invalid any expectation or data model that imposes a monotone relation between age and cancer (51, 45). As another example, if a trial was to be blinded but we knew the blinding was hopelessly compromised (e.g., by side effects), then we must adjust our expectations to account for the blinding failure. To get a sensitive diagnostic, however, we must not let our initial expectations be altered by the study results, because such alterations would bias the diagnostic toward detecting no discrepancy between expectations and observations.

Finally, expectations and interpretations need to take account of other variables in the model (155). For example, smoking is a well-known risk factor for heart disease; however, if the model includes major mediators of this smoking effect (such as blood pressure), the smoking coefficient may be diminished considerably relative to the expected total effect of smoking.

CURRENT MODELING STRATEGIES FOR CONFOUNDER CONTROL

Articles are usually clear about whether they modeled outcomes (as in risk or rate regression), exposure [as in propensity scoring (60, 119), E-estimation (118), and inverse-probability weighting

(60, 115)], or both [as in doubly robust methods (7, 60, 83, 99, 136, 140, 147)]. Some are also careful to exclude variables on causal grounds, such as instrumental variables and variables affected by exposure or disease, leaving only potential confounders for selection (24, 44, 72, 109, 121).

Beyond these basics, it may be difficult to tell if a prespecified modeling strategy was used. Often, however, we see variants of the following strategies, each of which may be applied to create outcome or exposure models:

1. Adjust all: Enter all the potential confounders in the model (only one set of covariates is considered, although the form of the model may be varied).
2. Predictor selection: Select covariates on the basis of some measure of their ability to predict outcome or exposure (or both) given other covariates in the model.
3. Change in estimate (CIE) selection: Select covariates on the basis of the change in the exposure effect estimate upon excluding them, given the other covariates in the model (14, 89, 125).

A problem with all these strategies is that they are not based on maximizing accuracy (minimizing bias and variance) in estimating target effects defined in an explicit causal model, and indeed they have no formal justification when there are many covariates to consider. As we now describe, each strategy has further problems.

Why Not Adjust for Every Available Covariate?

One study proposed that a relevant criterion for estimating causal effects is to adjust for all covariates known to cause exposure, disease, or both (145). This criterion can identify covariate subsets sufficient for confounding control, but it has some practical drawbacks. The set of covariates it identifies can be far larger than needed for adequate confounding control (far from minimally sufficient) and may be clumsy for subsequent analyses. Furthermore, even the largest studies can be small relative to the number of covariates potentially fulfilling this criterion, resulting in the breakdown of conventional fitting methods such as maximum likelihood (including its conditional and partial versions) (59, 76, 134). Finally, there may be many covariates whose causal status (and thus their fulfillment of the criterion) is uncertain.

Although sample-size adequacy for fitting methods is often judged by rules similar to "at least 10 subjects per regression coefficient," such rules take no account of exposure and disease frequencies and so can be too pessimistic when only an exposure coefficient is targeted (149) and too optimistic when some exposure–outcome combination is rare (76). In the latter case, controlling too many variables by conventional means can lead to or aggravate two closely related problems: (*a*) data sparsity, in which full control results in too few subjects at crucial combinations of the variables, with consequent inflation of estimates (59, 76, 116, 134), and (*b*) multicollinearity, by which we mean high multiple correlation (or more generally, high association) of the controlled variables with study exposures (116). In particular, if we include covariates that together are highly predictive of an exposure but are not all necessary to control confounding, the resulting effect estimate may be inflated or have unnecessarily wide confidence intervals (15, 26, 116). These problems increase as the ratio of number of covariates to sample size increases, motivating strategies to reduce the number of modeled covariates (116; S. Greenland & N. Pearce, unpublished manuscript, "Modeling Strategies for Observational Epidemiology").

Strategies Based on Predictive Modeling

Traditionally, variable selection is based solely on predicting observed outcomes under the observed distribution of exposure and covariates. The selection criterion is usually significance testing

of coefficients, as in conventional stepwise regression (37) [even though all modern software allows use of better criteria (60, 79, 80, 132)]. Stepwise regression (37) attempts to achieve parsimonious noncausal prediction, searching for a model that explains the most outcome variation with the fewest variables. The goal itself is reasonable for clinical prediction whenever obtaining variables incurs notable costs; for example, there are clear practical benefits if we can predict cardiovascular disease risk with negligible loss in accuracy by collecting information on 5 variables instead of 30. Even with this goal, however, ordinary stepwise regression has many flaws and has several more when used for estimating effects (6, 32, 38, 40–42, 47, 50, 53, 54, 79, 84, 127, 132, 133, 148, 152).

For covariates subject to selection, decisions about adding (or deleting) a covariate are made according to whether adding (or deleting) the covariate significantly improves (or reduces) the fit of the model, whereby "significantly" is usually defined by some arbitrary cutoff (α-level) for the coefficient p-value, usually 0.05, but preferably much higher (e.g., 0.20) for confounder selection (27, 47). Such criteria are equivalent to assessing whether the covariate explains a significant proportion of the residual variation (the outcome variation that remains given the preexisting variables in the model). They are also equivalent to using the p-value for testing the covariate when it is in the model.

Even for pure clinical prediction, ordinary stepwise regression and other significance-based selection procedures can give very distorted p-values and confidence intervals (2, 6, 21, 22, 32, 38, 40–42, 46, 79, 84, 86, 127, 132, 133). The general problem is that these traditional algorithms use no cross-validation and so do not account for preliminary testing (the same data being used to both fit and test the model). As a result, in outcome models, they produce p-values for the exposure effect that are too small (i.e., overstate significance) and confidence intervals that are too narrow (6, 32, 38, 40–42, 47, 50, 53, 54, 79, 84, 127, 132, 133, 148, 152). In addition, the resulting model often yields much poorer predictions than can be obtained with modern techniques (79, 80, 132). These defects are especially bad for fields (such as epidemiology) that are plagued by charges of generating too many false positives and inaccurate predictions. Defects can be corrected by using advanced resampling and cross-validation methods (80, 122, 132, 140), but these corrections remain uncommon in commercial software and thus are rare in published studies. Some of these problems can be moderated by replacing significance tests with selection criteria that penalize for model complexity, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) (79, 80, 132); nonetheless, the confidence intervals produced by these methods will remain too narrow (86).

Another persistent objection to outcome modeling is that it will select weak confounders or nonconfounders preferentially over strong confounders if the weak confounders are better than strong confounders at predicting the outcome (46). A parallel objection to exposure modeling is that selection based solely on predicting exposure will select weak confounders preferentially over strong confounders if the weak confounders are better than the strong confounders at predicting exposure (15, 31, 146, 154). Thus, covariates that statistically explain (are associated with) the most observed disease or exposure variation, or are most statistically significant (have the smallest p-value), need not be the same as the covariates that are most important in terms of confounding control (47, 71, 102) or public health importance (106).

These seemingly paradoxical facts arise because outcome and exposure models consider only one of the several parameters that determine confounding. In the extreme, an outcome risk factor may be strongly associated with the outcome, with a small p-value, and yet not be associated with exposure, whereas an exposure risk factor may be strongly associated with the exposure, with a small p-value, and yet not be associated with the outcome. Neither factor will be a confounder, yet both could be selected for adjustment in preference to actual confounders. For example, age is a strong risk factor for most diseases, but it will not be a confounder if it is independent of exposure

(as with outdoor air pollution within small enough areas). Conversely, a covariate may have a nonsignificant ($p > 0.05$) association with outcome or exposure owing to sample-size limitations, but it may still be an important confounder (71).

The problems just described often arise in pursuit of model goodness-of-fit; in particular, some variables may not be included in the model because they do not significantly improve the fit, even though they are important confounders. Global or omnibus tests of fit are especially inadequate for confounder selection because there can be many models that fit equally well but correspond to very different confounder effects and exposure effect estimates (116).

The preceding problems explain why ordinary model-fitting criteria such as significance testing of predictors have long been considered inappropriate for the assessment of confounding (14, 47, 71, 74, 89) [indeed, there are questions about whether tests can be justified for any epidemiologic purpose (48, 131)]. The same problems apply to exposure modeling, as in propensity scoring (PS) or inverse probability of treatment weighting (IPTW), but they manifest somewhat differently than they do in outcome modeling. In exposure modeling, the only association being tested is that of the potential confounder with exposure. This reorientation eliminates the downward bias of exposure–outcome *p*-values and standard errors produced by covariate deletion using the covariate–outcome *p*-values. Nonetheless, the preferential selection of variables that predict exposure will worsen multicollinearity problems, often increasing variance without sufficient compensation via bias reduction (15, 30, 31, 109, 110, 146). Furthermore, in case-control studies, exposure modeling can lead to bias (98, 136), which can be avoided by using special algorithms that use the outcome in the exposure model (136). Standard advice for mitigating this problem is to include only factors predictive of the outcome (as well as exposure) in the exposure model (4, 15, 109); this is sound advice but may still include far more covariates than desirable, especially in terms of precision loss relative to bias reduction.

Another objection to ordinary covariate selection criteria is that, when deciding whether to control for a variable, the null hypothesis is arguably not the correct one to test. If the goal is to avoid bias, it may be more accurate to start with the hypothesis that a particular variable is an important confounder and to decide to exclude the variable from the analysis only if the data indicate that it can be ignored (46, 47). For example, instead of testing whether a confounder coefficient is zero, one could test whether its absolute magnitude is larger than a given important size (a coefficient equivalence test) (81). More directly, one could test whether the change in the exposure effect estimate from controlling the covariate is larger than a certain size (a collapsibility equivalence test) (96, 101) as discussed below.

Change-in-Estimate Strategies

In contrast with traditional predictive modeling, CIE strategies select covariates on the basis of how much their control changes exposure effect estimates; this observed change is presumed to measure confounding by the covariate. Since the late 1970s, epidemiologic textbooks and articles have recommended CIE rather than significance testing of the covariate coefficient (14, 89, 125). Later versions (47, 74) suggest using change in the confidence limits instead because those are usually the final analysis product.

One caution to these approaches is that an accurate assessment of confounding may require examining changes from removing entire sets of covariates. Another caution, which arises if the disease frequency is high and measured by odds or rates, is that the change may partly reflect noncollapsibility of the effect measure rather than confounding (72, 104). Nonetheless, CIE methods have an advantage over selection based only on outcome or exposure prediction insofar as the selection criterion is on the scale used for contextual interpretation.

A complication of CIE methods is that one must choose an effect measure to judge change importance, where "importance" needs to be evaluated along a contextually meaningful scale. Except for linear models, this scale will not be that of model coefficients, but rather will be a nonlinear transformation of the coefficients. If the outcome is rare enough to ignore distinctions among risk, rate, and odds ratios, and the exposure effect is represented by a single coefficient β in a multiplicative model, one can simply compare the estimates of the constant (homogeneous) ratio effect $\exp(\beta)$ from models with and without the covariate at issue. Otherwise it may be necessary to compare standardized (marginal) risk ratios estimated from the models (104).

Suppose RR_a and RR_u denote the estimated risk ratio with and without adjustment for the covariate; then RR_a/RR_u is traditionally used to judge change importance. However, in public health applications in which total caseload is of primary concern, arguably the exposed attributable fraction (AF) $= (RR - 1)/RR$ is more relevant, in which case change could be measured by $AF_a - AF_u$, perhaps multiplied by 100 to express change in the percent caseload attributed to exposure. Many variations could arise depending on the ultimate target parameter.

Outcome Modeling, Exposure Modeling, or Both?

Outcome modeling is usually the simplest approach, especially when there is only one outcome but multiple exposures or an exposure with multiple levels. Nonetheless, exposure modeling may provide more valid effect estimates when there is more information for modeling exposures than for modeling outcomes. This advantage would usually arise when exposure is a treatment (and thus much is known about reasons for its use) or when the exposure is common but the outcome is not (and thus there are relatively few cases available for outcome modeling). Nonetheless, exposure modeling risks greater imprecision than outcome modeling does if the exposure is predicted better than necessary for confounding control (4).

Methods that model both outcome and exposure (including doubly robust methods) (7, 60, 83, 99, 136, 140, 147) avoid having to make a choice, but at the cost of more modeling effort. They have the option of using more data information with potential accuracy gains as a result. They may also use different covariates in the two models: Only confounders requiring the most accurate control (e.g., age) need appear in both models; minor or doubtful covariates may be limited to the model in which their role is better understood, knowing that their adequacy of control depends entirely on the accuracy of their specification in that model.

Parsimony versus Confounding

Both predictive modeling and CIE lack strong theoretical foundation for confounder selection and share several weaknesses. As commonly implemented, both assess confounding solely on the basis of the analysis data, ignoring the earlier data that led to consideration of certain variables as potential confounders. Both have parsimony as a key goal: to find a model that maximizes prediction or confounding control using few covariates. Nonetheless, parsimony is not a worthwhile goal in itself; its benefits must be demonstrated to justify its use as a criterion. Although parsimony can improve predictive accuracy and simplify analysis and presentation, typical arguments offered for parsimony are fallacious (87).

Some have argued that apparently weak confounders should be deleted from the model because "the use of a reduced model. . .can sometimes lead to a gain in precision" (89), pointing to the smaller estimated standard errors from smaller (simpler) fitted models as exhibiting that benefit. Unfortunately, this apparent variance reduction in single data sets is largely illusory because it

ignores the component of variance due to model selection. Again, there are ways to account for this problem (79, 80, 132, 140), but we know of none that are easy to implement with popular software.

An objection to all variable selection is that if our goal is to estimate the average effect of a particular exposure rather than all the model coefficients, then there is no direct reason for concern about accuracy or parsimony in estimating individual confounder effects. In that case, all that matters is whether the resulting model, taken as a whole, successfully removes confounding by the covariates. This argument has often been taken to imply that we should adjust for all measured potential confounders or at least a maximal number (47, 124, 145, 146), but again this view takes no account of the problems that can arise from doing so (62, 109, 110, 129, 130, 153).

Studies of Simple Strategies

There are many possible ways to use significance testing and CIE separately or together to reduce the number of model covariates. For the simplest methods, simulation studies (96, 101, 103) appear to confirm earlier suggestions (27) that false negatives (incorrect exclusion of confounders) are a greater threat to effect estimation accuracy than are false positives (incorrect inclusion of nonconfounders), consonant with theoretical criteria (145) and supporting weak exclusion criteria. As an illustration, consider one simple simulation study (96), which compared the following strategies:

I. Significance test the covariate coefficient in the outcome model; e.g., delete if the coefficient's p-value is under 0.05.

II. See whether the change in the exposure effect estimate from adjusting for the covariate falls outside an interval of practical equivalence; e.g., delete if $0.91 < RR_a/RR_u < 1.1$ (which is the 10%-change rule for the risk ratio modified to be proportionally symmetric).

III. Significance test the change from adjusting for the covariate (collapsibility testing); e.g., delete if a test of $RR_a/RR_u = 1$ yields $p < 0.05$.

IV. Test whether the change from adjusting for the covariate (noncollapsibility testing) is important (falls outside an interval of practical equivalence); e.g., delete if the 95% confidence interval for RR_a/RR_u (not just the point estimate) falls between 0.80 and 1.25 [this is a 0.05-level equivalence test (11, 81)].

V. A hybrid strategy uses a weighted average of $\ln(RR_a)$ and $\ln(RR_u)$ as the estimated exposure effect to reduce mean-squared error (49); this is not a selection strategy but instead a method for partial adjustment of all candidate covariates.

Strategies based on equivalence criteria (II, IV) performed best when the equivalence interval was narrow [e.g., $0.91 < RR_a/RR_c < 1.1$ for (II)], whereas significance-test strategies (I, III) performed best when the α-level was very high (e.g., using $p < 0.20$ instead of $p < 0.05$); thus, for all approaches, it appeared more important to avoid excluding possible important confounders than to avoid including weak confounders or nonconfounders.

None of the above selection strategies have a good grounding in theory, in that they are not derived to minimize error in any dimension; they are merely heuristics for finding simpler models that are not misleading for subsequent analysis and presentation. They thus leave considerable room for improvement. Noteworthy in this regard is the substantial literature in algorithmic modeling (machine learning), which has found that the performance of most simple predictive-modeling algorithms can be boosted considerably via computer-intensive methods, such as cross-validation or bootstrapping, to a level that rivals much more sophisticated algorithms (80). Such algorithms have been used to improve performance of exposure-modeling methods (95, 99, 105, 114, 126, 156) and can be applied similarly to outcome modeling (80) as long as exposure is forced into the model.

Further Considerations for Effect Estimation (Causal Inference)

All procedures require a target against which error (whether systematic or random) will be measured and a scale or loss function for measuring error and performance. Although there have been several applications of algorithmic exposure modeling for causal targets (140), the machine-learning algorithms in common software assume that the target is passive prediction to populations identical to that observed, apart from random variation. In outcome modeling, this amounts to success in passively predicting the outcome over the observed joint distribution of the exposure and covariates; in exposure modeling, this amounts to success in predicting the exposure over the observed joint distribution of covariates.

Effect estimation involves a different target: prediction of the outcome variable under at least two different exposure regimes and possibly under different covariate distributions than observed (65, 61, 83, 108, 140). Algorithms for passive prediction can be exploited for such causal prediction if care is taken to shift the final prediction target from passive to potential outcomes. Specifically, an individual's outcome under a treatment regime is only a potential outcome until that regime is chosen; after it is chosen, the other regimes become counterfactual for that individual, and outcomes under those regimes thus become unobservable. With this conceptualization, the problem of causal inference can be recast as a missing-data problem, opening the way for use of concepts and tools for missing-data analysis in causal inference (123, 140).

There are, however, elements of potential-outcome modeling that are not identified by statistical experiments, which have led to objections and alternatives (28, 29). Nonetheless, both potential-outcome models and their competitors must be elaborated considerably when exposure is ambiguously defined or is not an intervention or decision (e.g., age, sex) (56, 82, 144, 143). As argued by the discussants of Dawid (28), potential-outcome models are valuable if not indispensable for providing insights needed to develop and criticize causal questions and inference methods in these cases (56, 82, 83, 108).

SUMMARY AND CONCLUSIONS

Students are often taught that statistical significance testing is inappropriate for confounder evaluation and that CIE methods are preferable (14, 20, 60, 89, 125). Nonetheless, they are also often taught that the goal of modeling is to produce a model that is as simple as possible (parsimonious) while providing an adequate fit to the data, without recognizing that this goal is closer to that of passive prediction (as in ordinary stepwise regression) than it is to that of accurate estimation or prediction of effects. Confusion is exacerbated when the different goals of noncausal prediction and effect estimation are not made explicit because these differences lead to different modeling strategies. Further confusion arises from rigid application of idealized estimation criteria to messy observational settings in which the supporting theory is at best a thin and possibly misleading sketch of the real context.

We have emphasized that parsimony and goodness-of-fit are inappropriate end goals for modeling, as indicated by simulation studies in which full-model analysis sometimes outperforms conventional selection strategies (96, 101, 153). Conversely, however, rejecting any use of data-based model selection for causal inference is also inappropriate because it ignores the harsh reality that even databases of studies with hundreds of thousands of individuals often have limited numbers of pivotal observations (such as exposed cases). Coupled with the availability of what may be dozens or even hundreds of variables, some kind of targeted approach to error control is essential, accounting for both systematic and random errors. Taking these precautions will lead to restoration of parsimony as a heuristic, formalized in model-dimension (degree-of-freedom) reduction strategies such as shrinkage, algorithmic modeling, or combinations thereof (62, 85, 99, 158).

The simple strategies we have discussed are relevant to common borderline situations in which control of all potential confounders may be possible, yet benefits are expected from eliminating some or all variables whose inclusion is of uncertain value. The benefits may be theoretical, such as accuracy improvement, or practical, such as simplification of analysis or presentation. While we would prefer to see a shift toward methods with sound and relevant theoretical and practical foundations, the gap between state-of-the-art methodology and what is done in most publications has only grown despite early criticisms of common modeling strategies (94). This lag may be attributable in part to methodologic conservatism, which serves a purpose insofar as it takes time to evaluate new methodology in enough contexts to consider it adequately field tested. The accelerating pace of methodologic development might also be partly responsible.

Nonetheless, we also see major obstacles in the time and resource limits that constrain typical research teams. Our experience suggests that considerable software development and training will be needed to facilitate use of better methods before simple strategies can be retired. We discuss elsewhere some easily implemented methods with stronger theoretical justification and broader capabilities than the simple strategies reviewed above (59, 134). Regardless of modeling approach, however, we caution that no methodology is foolproof (especially when faced with uncontrolled confounding, selection bias, or measurement error) and that modeling methods should be documented in enough detail so that readers can interpret results in light of the strengths and weaknesses of those methods.

SUMMARY POINTS

1. Because models always fall far short of the complex reality under study, there are no best or optimal strategies for modeling. Strategies promoted as optimal or correct have these properties only under highly simplified models. When many covariates are available, however, traditional modeling strategies are demonstrably inferior to more modern techniques.
2. When we can discern a modeling strategy in epidemiological and medical studies, it is usually a variant of one of these approaches: (*a*) use of a model with all measured covariates a priori identified as potential confounders; (*b*) testing of covariate coefficients to eliminate some variables from a model predicting outcome or exposure (as in stepwise regression); or (*c*) using a change-in-estimate criterion to eliminate variables.
3. The goal of testing covariate coefficients is to obtain a model that predicts observed outcomes or exposures well with a minimal number of variables, whereas the goal of change-in-estimate criteria is to obtain a model that controls most or all confounding with a minimal number of variables.
4. The main goal of a statistical analysis of effects should be the production of the most accurate (least erroneous) effect estimates obtainable from the data and available software. None of the common strategies are targeted toward that goal, and better strategies are not yet integrated into popular commercial software.
5. Regardless of the modeling strategy chosen, it is important that authors document the strategy used so that readers can interpret the results in light of the strategy's strengths and weaknesses.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Steven Cole, Katherine J. Hoggatt, Mohammad A. Mansournia, and Charles Poole for helpful comments on earlier drafts of this paper. Neil Pearce also thanks Tony Blakely, Naomi Brewer, Tim Clayton, Simon Cousens, Rhian Daniel, Shah Ebrahim, Chris Frost, Michael Hills, Nancy Krieger, Dorothea Nitsch, George Davey Smith, Lorenzo Richiardi, Bianca de Stavola, and Jan Vandenbroucke for their valuable comments on earlier work leading to the present paper. The Centre for Public Health Research is supported by a program grant from the Health Research Council of New Zealand.

LITERATURE CITED

1. Ahrens W, Pigeot I, eds. 2014. *Handbook of Epidemiology*. New York: Springer. 2nd ed.
2. Arabatzis AA, Gregoire TG, Reynolds MR Jr. 1989. Conditional interval estimation of the mean following rejection of a two-sided test. *Commun. Stat.-Theory Methods* 18:4359–73
3. Austin PC. 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28:3083–107
4. Austin PC. 2014. A tutorial on the use of propensity score methods with survival time or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat. Med.* 33:1242–58
5. Austin PC, Brunner LJ. 2004. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Stat. Med.* 23:1159–78
6. Bancroft TA, Han CP. 1977. Inference based on conditional specification - note and a bibliography. *Int. Stat. Rev.* 45:117–27
7. Bang H, Robins JM. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61:962–72. Erratum. 2008. *Biometrics* 64(2):650
8. Bayarri MJ, Berger JO. 2004. The interplay of Bayesian and frequentist analysis. *Stat. Sci.* 19:58–80
9. Berger JO, Berry DA. 1988. Statistical analysis and the illusion of objectivity. *Am. Sci.* 76:159–65
10. Berzuini C, Dawid AP, Bernardinelli L, eds. 2012. *Causal Inference: Statistical Perspectives and Applications.* Chichester, UK: Wiley
11. Blackwelder WC. 1998. Equivalence trials. In *Encyclopedia of Biostatistics*, ed. P Armitage, T Colton, pp. 1367–72. New York: Wiley
12. Box GEP. 1980. Sampling and Bayes inference in scientific modeling and robustness. *J. R. Stat. Soc. Ser. A* 143:383–430
13. Box GEP. 1990. Comment. *Stat. Sci.* 5:390–91
14. Breslow NE, Day NE. 1980. *Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies.* Lyon: IARC
15. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. 2006. Variable selection for propensity score models. *Am. J. Epidemiol.* 163:1149–56
16. Carlin BP, Louis TA. 2000. *Bayes and Empirical Bayes Methods of Data Analysis*. New York: Chapman & Hall. 2nd ed.
17. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C. 2006. *Measurement Error in Non-Linear Models*. Boca Raton, FL: Chapman & Hall
18. Chatfield C. 1995. Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. Ser. A* 158:419–66
19. Chatfield C. 2002. Confessions of a pragmatic statistician. *J. R. Stat. Soc. Ser. D* 51:1–20

20. Checkoway H, Pearce N, Kriebel D. 2004. *Research Methods in Occupational Epidemiology*. New York: Oxford Univ. Press
21. Chiou P. 1997. Interval estimation of scale parameters following a pre-test for two exponential distributions. *Comput. Stat. Data Anal.* 23:477–89
22. Chiou P, Han CP. 1995. Conditional interval estimation of the exponential location parameter following rejection of a pretest. *Commun. Stat.-Theory Methods* 24:1481–92
23. Colditz GA, Manson JE, Hankinson SE. 1997. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J. Womens Health* 6:49–62
24. Cole SR, Platt RW, Schisterman EF, Chu HT, Westreich D, et al. 2010. Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* 39:417–20
25. Copas JB. 1983. Regression, prediction and shrinkage. *J. R. Stat. Soc. Ser. B* 45:311–54
26. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. 2011. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J. Clin. Epidemiol.* 64:1463–64
27. Dales LD, Ury HK. 1978. An improper use of statistical significance testing in studying covariables. *Int. J. Epidemiol.* 4:373–75
28. Dawid AP. 2000. Causal inference without counterfactuals. *J. Am. Stat. Assoc.* 95:407–48
29. Dawid AP. 2012. The decision-theoretic approach to causal inference. See Ref. 10, pp. 25–42
30. Day NE, Byar DP, Green SB. 1980. Overadjustment in case-control studies. *Am. J. Epidemiol.* 112:696–706
31. De Luna X, Waernbaum I, Richardson TS. 2011. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* 98:861–75
32. Draper NR, Guttman I, Lapczak L. 1979. Actual rejection levels in a certain stepwise test. *Commun. Stat. A* 8:99–105
33. Efron B. 1975. Biased versus unbiased estimation. *Adv. Math.* 16:259–77
34. Efron B. 2005. Bayesians, frequentists, and scientists. *J. Am. Stat. Assoc.* 100:1–5
35. Efron B, Morris C. 1973. Stein's estimation rule and its competitors—empirical Bayes approach. *J. Am. Stat. Assoc.* 68:117–30
36. Efron B, Morris C. 1975. Data analysis using Stein's estimator and its generalizations. *J. Am. Stat. Assoc.* 70:311–19
37. Efroymson MA. 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*, ed. A Ralston, HS Wilf, pp. 191–203. Hoboken, NJ: Wiley
38. Faraway JJ. 1992. On the cost of data analysis. *J. Comput. Graph. Stat.* 1:213–19
39. Feyerabend P. 1993. *Against Method*. New York: Verso
40. Flack VF, Chang PC. 1987. Frequency of selecting noise variables in subset regression analysis—a simulation study. *Am. Stat.* 41:84–86
41. Freedman DA. 1983. A note on screening regression equations. *Am. Stat.* 37:152–55
42. Freedman DA, Navidi W, Peters SC. 1988. On the impact of variable selection in fitting regression equations. In *On Model Uncertainty and Its Statistical Implications*, ed. TK Dijlestra, pp. 1–16. Berlin: Springer-Verlag
43. Gelman A, Shalizi CR. 2013. Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.* 66:8–38
44. Glymour MM, Greenland S. 2008. Causal diagrams. See Ref. 120, pp. 183–209
45. Greenland S. 1986. Partial and marginal matching in case-control studies. In *Modern Statistical Methods in Chronic Disease Epidemiology*, ed. SH Moolgavkar, RL Prentice, pp. 35–49. New York: Wiley
46. Greenland S. 1989. Comment: cautions in the use of preliminary test estimators. *Stat. Med.* 8:669–73
47. Greenland S. 1989. Modeling and variable selection in epidemiologic analysis. *Am. J. Public Health* 79:340–49
48. Greenland S. 1990. Randomization, statistics, and causal inference. *Epidemiology* 1:421–29
49. Greenland S. 1991. Reducing mean squared error in the analysis of stratified epidemiologic studies. *Biometrics* 47:773–75
50. Greenland S. 1993. Methods for epidemiologic analyses of multiple exposures: a review and a comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. *Stat. Med.* 12:717–36

51. Greenland S. 1997. Estimating relative risk functions in case-control studies using a nonparametric logistic regression. *Am. J. Epidemiol.* 146:883–84
52. Greenland S. 2000. Principles of multilevel modelling. *Int. J. Epidemiol.* 29:158–67
53. Greenland S. 2000. When should epidemiologic regressions use random coefficients? *Biometrics* 56:915–21
54. Greenland S. 2001. Putting background information about relative risks into conjugate prior distributions. *Biometrics* 57:663–70
55. Greenland S. 2002. A review of multilevel theory for ecologic analyses. *Stat. Med.* 21:389–95
56. Greenland S. 2005. Epidemiologic measures and policy formulation: lessons from potential outcomes. *Emerg. Themes Epidemiol.* 2:5
57. Greenland S. 2005. Multiple-bias modeling for observational studies (with discussion). *J. R. Stat. Soc. Ser. A* 168:267–308
58. Greenland S. 2006. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int. J. Epidemiol.* 35:765–75
59. Greenland S. 2007. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int. J. Epidemiol.* 36:195–202
60. Greenland S. 2008. Introduction to regression modeling. See Ref. 120, pp. 418–55
61. Greenland S. 2008. Introduction to regression models. See Ref. 120, pp. 381–417
62. Greenland S. 2008. Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *Am. J. Epidemiol.* 167:523–29
63. Greenland S. 2009. Dealing with uncertainty about investigator bias: Disclosure is informative. *J. Epidemiol. Community Health* 63:593–98
64. Greenland S. 2009. Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Stat. Sci.* 24:195–210
65. Greenland S. 2012. Causal inference as a prediction problem: assumptions, identification, and evidence synthesis. See Ref. 10, pp. 43–58
66. Greenland S. 2012. Intuitions, simulations, theorems: the role and limits of methodology. *Epidemiology* 23:440–42
67. Greenland S. 2014. Regression methods for epidemiological analysis. See Ref. 1, pp. 1087–159
68. Greenland S. 2014. Sensitivity analysis and bias analysis. See Ref. 1, pp. 685–706
69. Greenland S, Lash TL. 2008. Bias analysis. See Ref. 120, pp. 345–80
70. Greenland S, Maclure M, Schlesselman JJ, Poole C. 1991. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* 2:387–92
71. Greenland S, Neutra RR. 1980. Control of confounding in the assessment of medical technology. *Int. J. Epidemiol.* 9:361–67
72. Greenland S, Pearl J. 2011. Adjustments and their consequences—collapsibility analysis using graphical models. *Int. Stat. Rev.* 79:401–26
73. Greenland S, Rothman KJ. 2008. Fundamentals of epidemiologic data analysis. See Ref. 120, pp. 213–37
74. Greenland S, Rothman KJ. 2008. Introduction to stratified analysis. See Ref. 120, pp. 258–82
75. Greenland S, Schlesselman JJ, Criqui MH. 1986. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am. J. Epidemiol.* 123:203–8
76. Greenland S, Schwartzbaum JA, Finkle WD. 2000. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am. J. Epidemiol.* 151:531–39
77. Gustafson P, Greenland S. 2009. Interval estimation for messy observational data. *Stat. Sci.* 24:328–42
78. Gustafson P, McCandless LC. 2010. Probabilistic approaches to better quantifying the results of epidemiologic studies. *Int. J. Environ. Res. Public Health* 7:1520–39
79. Harrell F. 2001. *Regression Modeling Strategies*. New York: Springer
80. Hastie T, Tibshirani R, Friendman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer. 2nd ed.
81. Hauck WW, Anderson S. 1986. A proposal for interpreting and reporting negative studies. *Stat. Med.* 5:203–9
82. Hernán MA. 2005. Invited commentary: Hypothetical interventions to define causal effects—afterthought or prerequisite? *Am. J. Epidemiol.* 162:618–20

83. Hernán MA, Robins JM. 2015. *Causal Inference*. New York: Chapman & Hall/CRC. In press
84. Hurvich CM, Tsai CL. 1990. The impact of model selection on inference in linear regression. *Am. Stat.* 44:214–17
85. Joffe MM. 2009. Exhaustion, automation, theory, and confounding. *Epidemiology* 20:523–24
86. Kabaila P. 2009. The coverage properties of confidence regions after model selection. *Int. Stat. Rev.* 77:405–14
87. Kelly KT. 2011. Simplicity, truth, and probability. In *Handbook of the Philosophy of Science: Philosophy of Statistics*, ed. PS Bandyopadhyay, MR Forster, pp. 983–1026. North Holland: Elsevier
88. Kendall MG. 1963. Comment on JO Irwin. The place of mathematics in medical and biological statistics. *J. R. Stat. Soc. Ser. A* 126:1–45
89. Kleinbaum D, Kupper LL, Morgenstern H. 1982. *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, CA: Lifetime Learn.
90. Krieger N. 2011. *Epidemiology and the People's Health: Theory and Context*. New York: Oxford Univ. Press
91. Lash TL. 2007. Heuristic thinking and inference from observational epidemiology. *Epidemiology* 18:67–72
92. Lash TL, Fox MP, Fink AK. 2009. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Boston: Springer
93. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. 2014. Good practices for quantitative bias analysis. *Int. J. Epidemiol.* 43:1969–85
94. Leamer E. 1978. *Specification Searches*. New York: Wiley
95. Lee BK, Lessler J, Stuart EA. 2010. Improving propensity score weighting using machine learning. *Stat. Med.* 29:337–46
96. Maldonado G, Greenland S. 1993. Simulation study of confounder-selection strategies. *Am. J. Epidemiol.* 138:923–36
97. Mallows CL, Nair VN. 1987. A unique unbiased estimator with an interesting property. *Am. Stat.* 41:205–6
98. Mansson R, Joffe MM, Sun WG, Hennessy S. 2007. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am. J. Epidemiol.* 166:332–39
99. McCaffrey DF, Ridgeway G, Morral AR. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* 9:403–25
100. McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. New York: Chapman & Hall
101. Mickey RM, Greenland S. 1989. The impact of confounder selection criteria on effect estimation. *Am. J. Epidemiol.* 129:125–37
102. Miettinen OS. 1976. Stratification by a multivariate confounder score. *Am. J. Epidemiol.* 104:609–20
103. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, et al. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. Epidemiol.* 174:1223–27
104. Pang M, Kaufman JS, Platt RW. 2014. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Stat. Methods Med. Res.* In press
105. Patorno E, Glynn RJ, Hernandez-Diaz S, Liu J, Schneeweiss S. 2014. Studies with many covariates and few outcomes selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology* 25:268–78
106. Pearce N. 2011. Epidemiology in a changing world: variation, causation and ubiquitous risk factors. *Int. J. Epidemiol.* 40:503–12
107. Pearce SC. 1983. The monstrous regiment of mathematicians. *Statistician* 32:375–78
108. Pearl J. 2009. *Causality: Models, Reasoning, and Inference*. New York: Cambridge Univ. Press. 2nd ed.
109. Pearl J. 2010. On a class of bias-amplifying variables that endanger effect estimates. In *Proc. 26th Conf. Uncertain. Artificial Intell.*, ed. P Grunwald, P Spirtes, pp. 417–24. Corvallis, OR: AUAI
110. Pearl J. 2012. Understanding bias amplification. *Am. J. Epidemiol.* 174:1223–27
111. Perlman MD, Wu L. 1999. The emperor's new tests (with discussion). *Stat. Sci.* 14:355–81
112. Phillips CV. 2004. Publication bias *in situ*. *BMC Med. Res. Methodol.* 4:20
113. Popper KR. 1959. *The Logic of Scientific Discovery*. New York: Basic Books
114. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. 2011. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am. J. Epidemiol.* 173:1404–13

115. Robins JM. 1999. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, ed. ME Halloran, D Berry, 116:95–134. New York: Springer
116. Robins JM, Greenland S. 1986. The role of model selection in causal inference from nonexperimental data. *Am. J. Epidemiol.* 123:392–402
117. Robins JM, Hernán MA. 2009. Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, ed. G Fitzmaurice, M Davidian, G Verbeke, G Molenberghs, pp. 553–600. New York: Chapman & Hall/CRC Press
118. Robins JM, Mark SD, Newey WK. 1992. Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* 48:479–95
119. Rosenbaum PR. 2002. *Observational Studies*. New York: Springer
120. Rothman KJ, Greenland S, Lash TL, eds. 2008. *Modern Epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins. 3rd ed.
121. Rothman KJ, Greenland S, Lash TL. 2008. Validity in epidemiologic studies. See Ref. 120, pp. 128–47
122. Royston P, Sauerbrei W. 2008. *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Chichester, UK: Wiley
123. Rubin DB. 1991. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 47:1213–34
124. Rubin DB. 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* 26:20–36
125. Schlesselman JJ. 1982. *Case-Control Studies: Design, Conduct, Analysis*. New York: Oxford Univ. Press
126. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. 2009. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20:512–22
127. Sclove SL, Morris C, Radhakrishnan R. 1972. Non-optimality of preliminary-test estimators for mean of a multivariate normal distribution. *Ann. Math. Stat.* 43:1481–90
128. Senn SJ. 1998. Mathematics: governess or handmaiden? *Statistician* 47:251–59
129. Shrier I. 2009. Propensity scores. *Stat. Med.* 28:1317–18
130. Sjölander A. 2009. Propensity scores and M-structures. *Stat. Med.* 28:1416–20
131. Sterne JAC, Smith GD. 2001. Sifting the evidence—what's wrong with significance tests? *BMJ* 322:226–31
132. Steyerberg EW. 2008. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer
133. Steyerberg EW, Eijkemans MJC, Habbema JDF. 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J. Clin. Epidemiol.* 52:935–42
134. Sullivan S, Greenland S. 2013. Bayesian regression in SAS software. *Int. J. Epidemiol.* 42:308–17
135. Susser M. 1977. Judgement and causal inference. *Am. J. Epidemiol.* 105:1–15
136. Tchetgen Tchetgen EJ. 2013. On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. *Am. J. Epidemiol.* 177:1314–16
137. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. 2009. Bias modelling in evidence synthesis. *J. R. Stat. Soc. Ser. A* 172:21–47
138. Valeri L, VanderWeele TJ. 2013. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods* 18:137–50
139. Vandenbroucke JP. 1987. Should we abandon statistical modeling altogether? *Am. J. Epidemiol.* 126:10–13
140. van der Laan MJ, Rose R. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer
141. VanderWeele TJ. 2009. On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology* 20:496–99
142. VanderWeele TJ. 2011. Subtleties of explanatory language: What is meant by "mediation"? *Eur. J. Epidemiol.* 26:343–46

143. VanderWeele TJ, Hernán MA. 2012. Causal effects and natural laws: towards a conceptualization of causal counterfactuals for nonmanipulable exposures with application to the effects of race and sex. See Ref. 10, pp. 101–13
144. VanderWeele TJ, Hernán M. 2013. Causal inference under multiple versions of treatment. *J. Causal Inference* 1:1–20
145. VanderWeele TJ, Shpitser I. 2011. A new criterion for confounder selection. *Biometrics* 67:1406–13
146. Vansteelandt S, Bekaert M, Claeskens G. 2012. On model selection and model misspecification in causal inference. *Stat. Methods Med. Res.* 21:7–30
147. Vansteelandt S, Daniel RM. 2014. On regression adjustment for the propensity score. *Stat. Med.* 33:4053–72
148. Viallefont V, Raftery AE, Richardson S. 2001. Variable selection and Bayesian model averaging in case-control studies. *Stat. Med.* 20:3215–30
149. Vittinghoff E, McCulloch CE. 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am. J. Epidemiol.* 165:710–18
150. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, et al. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 370:1453–57
151. Wakefield J, Salway R. 2001. A statistical framework for ecological and aggregate studies. *J. R. Stat. Soc. Ser. A* 164:119–37
152. Weiss RE. 1995. The influence of variable selection—a Bayesian diagnostic perspective. *J. Am. Stat. Assoc.* 90:619–25
153. Weng HY, Hsueh YH, Messam LLM, Hertz-Picciotto I. 2009. Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. *Am. J. Epidemiol.* 169:1182–90
154. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stuermer T. 2011. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol. Drug Saf.* 20:317–20
155. Westreich D, Greenland S. 2013. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am. J. Epidemiol.* 177:292–98
156. Westreich D, Lessler J, Funk MJ. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* 63:826–33
157. Whitehead AN. 1929. *Process and Reality*. New York: Harper
158. Wilson A, Reich BJ. 2014. Confounder selection via penalized credible regions. *Biometrics* 70:852–61
159. Ziman J. 1978. *Reliable Knowledge*. New York: Cambridge

Annual Review of Public Health
Volume 36, 2015

Contents

Symposium: Strategies to Prevent Gun Violence

Commentary: Evidence to Guide Gun Violence Prevention in America
Daniel W. Webster 1

The Epidemiology of Firearm Violence in the Twenty-First Century United States
Garen J. Wintemute 5

Effects of Policies Designed to Keep Firearms from High-Risk Individuals
Daniel W. Webster and Garen J. Wintemute 21

Cure Violence: A Public Health Model to Reduce Gun Violence
Jeffrey A. Butts, Caterina Gouvis Roman, Lindsay Bostwick, and Jeremy R. Porter 39

Focused Deterrence and the Prevention of Violent Gun Injuries: Practice, Theoretical Principles, and Scientific Evidence
Anthony A. Braga and David L. Weisburd 55

Epidemiology and Biostatistics

Has Epidemiology Become Infatuated With Methods? A Historical Perspective on the Place of Methods During the Classical (1945–1965) Phase of Epidemiology
Alfredo Morabia 69

Statistical Foundations for Model-Based Adjustments
Sander Greenland and Neil Pearce 89

The Elusiveness of Population-Wide High Blood Pressure Control
Paul K. Whelton 109

The Epidemiology of Firearm Violence in the Twenty-First Century United States
Garen J. Wintemute 5

Focused Deterrence and the Prevention of Violent Gun Injuries: Practice, Theoretical Principles, and Scientific Evidence
Anthony A. Braga and David L. Weisburd 55

Unintentional Home Injuries Across the Life Span: Problems and Solutions
Andrea C. Gielen, Eileen M. McDonald, and Wendy Shields 231

Sleep as a Potential Fundamental Contributor to Disparities in Cardiovascular Health
Chandra L. Jackson, Susan Redline, and Karen M. Emmons 417

Translating Evidence into Population Health Improvement: Strategies and Barriers
Steven H. Woolf, Jason Q. Purnell, Sarah M. Simon, Emily B. Zimmerman, Gabriela J. Camberos, Amber Haley, and Robert P. Fields 463

Environmental and Occupational Health

Fitness of the US Workforce
Nicolaas P. Pronk 131

Food System Policy, Public Health, and Human Rights in the United States
Kerry L. Shannon, Brent F. Kim, Shawn E. McKenzie, and Robert S. Lawrence 151

Regulating Chemicals: Law, Science, and the Unbearable Burdens of Regulation
Ellen K. Silbergeld, Daniele Mandrioli, and Carl F. Cranor 175

The Haves, the Have-Nots, and the Health of Everyone: The Relationship Between Social Inequality and Environmental Quality
Lara Cushing, Rachel Morello-Frosch, Madeline Wander, and Manuel Pastor 193

The Impact of Toxins on the Developing Brain
Bruce P. Lanphear 211

Unintentional Home Injuries Across the Life Span: Problems and Solutions
Andrea C. Gielen, Eileen M. McDonald, and Wendy Shields 231

Public Health Practice

Cross-Sector Partnerships and Public Health: Challenges and Opportunities for Addressing Obesity and Noncommunicable Diseases Through Engagement with the Private Sector
Lee M. Johnston and Diane T. Finegood 255

Deciphering the Imperative: Translating Public Health Quality Improvement into Organizational Performance Management Gains
Leslie M. Beitsch, Valerie A. Yeager, and John Moran 273

Identifying the Effects of Environmental and Policy Change Interventions on Healthy Eating
Deborah J. Bowen, Wendy E. Barrington, and Shirley A.A. Beresford 289

Lessons from Complex Interventions to Improve Health
Penelope Hawe 307

Trade Policy and Public Health
Sharon Friel, Libby Hattersley, and Ruth Townsend 325

Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health
Guthrie S. Birkhead, Michael Klompas, and Nirav R. Shah 345

What Is Health Resilience and How Can We Build It?
Katharine Wulff, Darrin Donato, and Nicole Lurie 361

Effects of Policies Designed to Keep Firearms from High-Risk Individuals
Daniel W. Webster and Garen J. Wintemute 21

Cure Violence: A Public Health Model to Reduce Gun Violence
Jeffrey A. Butts, Caterina Gouvis Roman, Lindsay Bostwick, and Jeremy R. Porter 39

Focused Deterrence and the Prevention of Violent Gun Injuries: Practice, Theoretical Principles, and Scientific Evidence
Anthony A. Braga and David L. Weisburd 55

Regulating Chemicals: Law, Science, and the Unbearable Burdens of Regulation
Ellen K. Silbergeld, Daniele Mandrioli, and Carl F. Cranor 175

The Response of the US Centers for Disease Control and Prevention to the Obesity Epidemic
William H. Dietz 575

Social Environment and Behavior

Immigration as a Social Determinant of Health
Heide Castañeda, Seth M. Holmes, Daniel S. Madrigal, Maria-Elena DeTrinidad Young, Naomi Beyeler, and James Quesada 375

Mobile Text Messaging for Health: A Systematic Review of Reviews
Amanda K. Hall, Heather Cole-Lewis, and Jay M. Bernhardt 393

Sleep as a Potential Fundamental Contributor to Disparities in Cardiovascular Health
Chandra L. Jackson, Susan Redline, and Karen M. Emmons 417

Stress and Type 2 Diabetes: A Review of How Stress Contributes to the Development of Type 2 Diabetes
Shona J. Kelly and Mubarak Ismail 441

Translating Evidence into Population Health Improvement: Strategies and Barriers
Steven H. Woolf, Jason Q. Purnell, Sarah M. Simon, Emily B. Zimmerman, Gabriela J. Camberos, Amber Haley, and Robert P. Fields 463

Using New Technologies to Improve the Prevention and Management of Chronic Conditions in Populations
Brian Oldenburg, C. Barr Taylor, Adrienne O'Neil, Fiona Cocker, and Linda D. Cameron 483

Commentary: Evidence to Guide Gun Violence Prevention in America
Daniel W. Webster 1

The Haves, the Have-Nots, and the Health of Everyone: The Relationship Between Social Inequality and Environmental Quality
Lara Cushing, Rachel Morello-Frosch, Madeline Wander, and Manuel Pastor 193

Cross-Sector Partnerships and Public Health: Challenges and Opportunities for Addressing Obesity and Noncommunicable Diseases Through Engagement with the Private Sector
Lee M. Johnston and Diane T. Finegood 255

Lessons from Complex Interventions to Improve Health
Penelope Hawe 307

What Is Health Resilience and How Can We Build It?
Katharine Wulff, Darrin Donato, and Nicole Lurie 361

Health Services

Assessing and Changing Organizational Social Contexts for Effective Mental Health Services
Charles Glisson and Nathaniel J. Williams 507

Policy Dilemmas in Latino Health Care and Implementation of the Affordable Care Act
Alexander N. Ortega, Hector P. Rodriguez, and Arturo Vargas Bustamante 525

Tax-Exempt Hospitals and Community Benefit: New Directions in Policy and Practice
Daniel B. Rubin, Simone R. Singh, and Gary J. Young 545

The Prescription Opioid and Heroin Crisis: A Public Health Approach to an Epidemic of Addiction
Andrew Kolodny, David T. Courtwright, Catherine S. Hwang, Peter Kreiner, John L. Eadie, Thomas W. Clark, and G. Caleb Alexander 559

The Response of the US Centers for Disease Control and Prevention to the Obesity Epidemic
William H. Dietz 575

Mobile Text Messaging for Health: A Systematic Review of Reviews
Amanda K. Hall, Heather Cole-Lewis, and Jay M. Bernhardt 393

Using New Technologies to Improve the Prevention and Management of Chronic Conditions in Populations
Brian Oldenburg, C. Barr Taylor, Adrienne O'Neil, Fiona Cocker, and Linda D. Cameron 483

Indexes

Cumulative Index of Contributing Authors, Volumes 27–36 597

Cumulative Index of Article Titles, Volumes 27–36 603

Errata

An online log of corrections to *Annual Review of Public Health* articles may be found at http://www.annualreviews.org/errata/publhealth