EDITORIAL

A Source of False Findings in Published Research Studies
Adjusting for Covariates

Helena Chmura Kraemer, PhD

Concern about erroneous conclusions of many published research findings has led to the conclusion that most published research findings are wrong.[1,2] What can be done about that? In what follows, I will focus on one common source of false findings: adjusting for covariates.

Here, *adjusting* means allowing variables to vary as they will but then using a mathematical model to assess their influence on the outcome. In contrast, *to control* means manipulation of variables by the researcher for a particular purpose (eg, in experimental design). Unfortunately, the terms *adjust* and *control* are often used as if they were synonymous. Adjusting often leads to false conclusions because the models used may not correspond to reality.

To illustrate this point, consider a randomized clinical trial (RCT) in which those sampled from the population of interest are randomly assigned to 2 treatment groups, T1 and T2. A valid test simply comparing the outcomes in the 2 groups tests the overall effect size (overall ES) that a randomly sampled patient from T1 has an outcome clinically preferable to that of a randomly sampled patient from T2.[3]

Often, the first table of an RCT report compares the baseline characteristics of the T1 vs T2 samples to assess the success of randomization, ignoring the fact that randomization (1) is a process, not an outcome, and (2) is meant to generate 2 random samples from the same population, not 2 matched samples. When a few baseline variables significantly differentiate the 2 groups at the 5% level, researchers often propose to adjust for those covariates in testing the treatment effect. This is post hoc testing (like offering to bet at prerace odds on a horse as it approaches the finish line), which frequently leads to false-positive results.

Any covariates to be used in adjusting should be specified a priori, listed in the RCT registration, and taken into consideration in the power analysis. Such adjustment changes the hypothesis to be tested from comparing all T1 patients vs all T2 patients (overall ES) to comparing T1 patients only with T2 patients matched in one way or another on the particular covariates proposed. Let's say that covariate ES is the ES for patients with one particular configuration of the covariates and typical ES is the ES specifically for patients who are at the mean of each such covariate (ie, for the typical patient). Overall ES, typical ES, and all possible covariate ES are the same only if the covariates are irrelevant to the treatment outcome. If the covariates are irrelevant, adjusting for those simply leads to a loss of power. If the covariates are not irrelevant, then estimation and testing of overall ES, typical ES, and covariate ES provide answers to different research questions.

The linear model used for covariate adjusting (eg, analysis of covariance [ANCOVA]) assumes, for all possible values of the covariates, that covariate ES is equal to typical ES; that is, that there is no interaction between the covariates and the treatment effect. If this assumption is violated, then the interactions that exist in the population (but are not included in the model) can bias the statistical tests and estimation of the treatment ES. Furthermore, not finding statistically significant interactions in the sample does not prove the null hypothesis that they do not exist in the population. Given these risks for bias, ANCOVA should not generally be used for such adjustment.

When treatment interactions are included in a linear model, how the variables are coded can impact the results.[4] The treatment effect refers to the treatment effect for patients having the zero value of all included covariates. Thus, if T1 and T2 were 2 treatments for Alzheimer disease, and the single covariate were chronological age at disease onset, the treatment effect would be the effect of the treatment for individuals with Alzheimer disease diagnosed as having the illness at age 0 years, which is a ludicrous result. Instead, age is better coded as deviations from the mean age at onset (centering at the mean). Then the treatment effect is typical ES and the interaction effect reflects the change in covariate ES as the covariate value changes. Examination of the covariate ES may well indicate to clinicians which patients will respond better to T1 or T2.[5] With multiple covariates, if each is centered at its mean,[4] the treatment effect tested is typical ES, the treatment effect for those at the mean of every covariate, which is sometimes a very small subpopulation.

There are still additional problems. For example, when multiple covariates are included, omitting interactions between them can introduce bias to the estimation of treatment ES. However, to include all interactions involving *m* covariates in a linear model requires estimation of 2^{m+1} parameters. To make matters even worse, the advantage of an RCT with random assignment is that, over replications, treatment choice and each covariate are uncorrelated. However, covariates may be correlated with each other (collinearity). Correlated variables share information. In fitting a model to the data, the computer is instructed to allocate the information shared between 2 variables to one variable or the other. The computer does this using information from within the sample. Because such information will change from one sample to another, the estimates of the adjusted treatment effects (typical ES and covariate ES) are unstable and difficult to replicate.

The bottom line is that covariates proposed a priori should always have strong rationale and justification and should be as few in number and as noncorrelated as possible. Often the best choice is to ignore covariates and to test and estimate overall ES and then to explore possible moderators of treatment response (ie, baseline variables for which covariate ES differs for different covariate values).[5] Subsequent hypothesis-testing studies can focus on those particular covariates that moderate treatment response.

Clearly, researchers bear the primary responsibility for the veracity of the findings they report. Post hoc hypothesis testing should not lead to conclusions. Instead, hypothesis-generating (exploratory) studies on the same data can provide rationale and justification for future hypothesis-testing studies. An interpretable ES and its confidence interval should be presented with each *P* value.[6,7] There should be no surprise when some statistically significant results (even with $P = 10^{-10}$) are of no clinical or practical significance. Tests that are not statistically significant should be regarded as indicative of poorly justified, designed, or executed hypothesis-testing studies, not as proof of the null hypothesis. Knowing and checking the assumptions made in any model is essential (eg, absence of interactions in ANCOVA models) and a clear interpretation of each parameter tested or estimated (such as overall ES vs typical ES vs covariate ES in an RCT) should be presented.

Reviewers and editors provide an additional level of protection against false findings in the literature. They should be sensitive to the problems of post hoc testing and refrain from suggesting post hoc hypotheses, such as inclusion of covariates or outcomes the researchers had not considered a priori. They should insist on ESs and confidence intervals that can be interpreted by the intended readers of the report. Finally, they should be alert to statistical errors justified by "But that's the way everyone does it!" (eg, ANCOVA to adjust for multiple, interacting, and collinear variables).

We cannot eliminate false findings with such efforts; however, we can get the percentage of false findings closer to the conventional 5%.

ARTICLE INFORMATION

Author Affiliations: Department of Psychiatry and Behavioral Sciences, Stanford University (Emerita), Palo Alto, California; Statistical Editor, *JAMA Psychiatry*.

Corresponding Author: Helena Chmura Kraemer, PhD, Department of Psychiatry and Behavioral Sciences, Stanford University (Emerita), 1116 Forest Ave, Palo Alto, CA 94301 (hckhome@pacbell.net).

Published Online: August 5, 2015. doi:10.1001/jamapsychiatry.2015.1178.

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294(2):218-228.

2. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

3. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry*. 2006;59(11):990-996.

4. Kraemer HC, Blasey CM. Centring in regression analyses: a strategy to prevent errors in statistical inference. *Int J Methods Psychiatr Res*. 2004;13(3):141-151.

5. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. *JAMA*. 2006;296(10):1286-1289.

6. Grissom RJ, Kim JJ. *Effect Sizes for Research: Univariate and Multivariate Applications.* New York, NY: Routledge; 2012.

7. Cumming G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis.* New York, NY: Routledge; 2012.

Low Resting Heart Rate as an Unequivocal Risk Factor for Both the Perpetration of and Exposure to Violence

Adrian Raine, DPhil

Low resting heart rate (RHR) has for some time been suspected to be a risk factor for crime and violence. One prior meta-analysis of 40 studies with a combined sample of 5868 individuals documented an association between low RHR and high antisocial and aggressive behavior in child and adolescent samples.[1] But questions remain. Is low RHR also a robust risk factor for adult violence? Does it prospectively predict violence? And what about less serious offenses such as traffic violations?

Related article page 971

In this issue of *JAMA Psychiatry*, in an exceptional study based on data on 710 264 Swedish men, Latvala and colleagues[2] document that low RHR at age 18 years predicts adult violence more than 30 years later. With a mean follow-up of 18.1 years, low RHR resulted in a 49% increased risk for violent crime after taking account of multiple confounding factors—a powerful confirmation of earlier studies. Their sample size, which was more than 100 times larger than all prior combined samples, places the empirical basis of their longitudinal findings beyond further dispute. Furthermore, they move this literature into new territory, documenting that low RHR predicts severe violence, less-severe violence, drug-related crime, property crime, and even traffic crime. This evidence establishes low RHR as a marker for broad rule-breaking behavior in general, although it particularly predicts serious violence.