Received: 4 January 2021 | Accepted: 3 February 2021

DOI: 10.1111/biom.13493

Biometrics WILEY

DISCUSSION

Discussion on "Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment for binary, ordinal, and time-to-event outcomes"

Michael A. Proschan

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases

Correspondence
Michael A. Proschan, Biostatistics Research Branch, NIAID, 5601 Fishers Lane, Rm 4C30, MSC 9820, Rockville, MD 20852.
Email: proscham@niaid.nih.gov

Abstract

Benkeser *et al.* present a very informative paper evaluating the efficiency gains of covariate adjustment in settings with binary, ordinal, and time-to-event outcomes. The adjustment method focuses on estimating the marginal treatment effect averaged over the covariate distribution in both arms combined. The authors show that covariate adjustment can achieve power gains that could find answers more quickly. The suggested approach is an important weapon in the armamentarium against epidemics like COVID-19. I recommend evaluating the procedure against more traditional approaches for conditional analyses (e.g., logistic regression) and against blinded methods of building prediction models followed by randomization-based inference.

KEYWORDS

binary, ordinal, and survival endpoints, covariate adjustment, marginal effects, randomization tests

Congratulations to Benkeser *et al.* (2020) for their excellent and timely work to improve the statistical analysis of COVID-19 trials with binary, ordinal, and time-to-event outcomes by incorporating covariate information. The methods they describe are under-utilized and not well appreciated by the clinical trial community. Picking appropriate endpoints and analysis methods can be very challenging in the context of a pandemic (Dodd *et al.*, 2020). On the one hand, we want simple, robust methods that will provide unequivocal evidence about a treatment. On the other hand, we want an answer as quickly as possible. The latter consideration argues for taking full advantage of covariate information to reduce sample size.

The marginal approach to covariate adjustment evaluated in Benkeser *et al.* is different from the conditional approach often used in clinical trials. For example, suppose that Y is binary. The traditional covariate adjustment method uses logistic regression to model the conditional probability $P(Y = 1 \mid A, \mathbf{X})$, and then makes inferences about the treatment effect using the coefficient for the arm variable A. Correct conclusions are contingent on correctness of this conditional model. Logistic regression is only the first step for Benkeser *et al.* They use it to estimate $P(Y = 1 \mid A = 0, \mathbf{X})$ and $P(Y = 1 \mid A = 1, \mathbf{X})$. They compute arm-specific marginal event probabilities $P(Y = 1 \mid A = a)$ by averaging $P(Y = 1 \mid A = a, \mathbf{X})$ over the covariate distribution of all patients (combined across arms). The treatment effect estimate is some function of these arm-specific marginal probabilities (e.g., the risk difference, relative risk, or odds ratio). Irrespective of the correctness of the logistic regression model, the marginal estimator provides a valid measure of treatment effect.

Benkeser *et al.* adopt a similar marginal approach for ordinal and time-to-event outcomes. A model is used to compute the conditional cumulative distribution function (CDF) given covariates, and arm-specific marginal CDFs are estimated by averaging over the covariate distribution combined across arms. The treatment effect estimate is

some function of the marginal CDFs. For ordinal (resp., time-to-event) outcomes, the treatment effect estimate is the difference in means, Mann–Whitney estimand, or log odds ratio (resp., the difference in restricted mean survival time or survival probability at a specific time, or relative risk at a specific time).

An important aspect of the proposed approach is that the initial conditional model allows separate coefficient values in the two arms. A similar idea has been used with missing data; imputation models may use separate coefficients in the different arms to increase the ability to predict the value of the missing outcome. Clinical trialists view this aspect with skepticism because of the perception that the type I error rate might be inflated. Nonetheless, asymptotic arguments and simulation results of Benkeser *et al.* show that tests and confidence intervals for this marginal estimator have correct error rates and achieve efficiency gains relative to unadjusted estimates of treatment effect. Using separate coefficients in the two arms means there will be a nonzero treatment by covariate interaction estimate. In most randomized trials, dramatic between-arm differences in true slopes are unlikely, begging the question of how much would be lost in assuming equal slopes. The traditional view is that a model with interaction makes interpretation difficult. The marginal approach posits that the average effect over the observed distribution of covariate values remains a meaningful summary. Still, the marginal approach that uses different slopes may complicate the combining of treatment effects across trials or even across time in the same trial. For example, the authors discuss group-sequential monitoring. The distribution of covariate values might change over the course of the trial. Earlier patients might be sicker or less sick than later patients. The estimand from the marginal approach may not be estimating the same thing over time. Of course, the treatment effect could change over time in a clinical trial regardless of the analysis method, but it seems worse when the sole reason for a changing treatment effect is that the distribution of baseline covariates is changing. With the conditional approach with no interactions, the treatment effect (e.g., adjusted odds ratio) is assumed to be the same for different covariate values. If true, this assumption facilitates synthesis of results over time, across subgroups, or across trials.

A different approach proposed by Gail *et al.* (1988) for testing can also be used to construct confidence intervals, at least in the binary outcome case. Using all patients from both arms and without knowledge of treatment assignments, build a regression model using any technique—examination of diagnostic plots, stepwise regression, and so on. Let $\hat{Y}_i$ be the predicted value of Y_i using the final regression model, and let $R_i = Y_i - \hat{Y}_i$ be the ith residual. Compute the observed treatment effect as $\hat{\delta} = \bar{R}_T - \bar{R}_C$, the difference in average residuals between treatment and control arms. Compute a p-value using a randomization test. For binary endpoints, one can view the estimand as the covariate-adjusted number of events saved per person treated. A confidence interval can also be computed by inverting the randomization test using the potential outcomes approach as described by Wang and Rosenberger (2020). An important advantage is that randomization-based methods can be used even with covariate-adaptive randomization (see Simon and Simon, 2011). In contrast, Benkeser *et al.* require randomization to be independent of covariate values (although they later point out that it can be modified to accommodate stratified randomization).

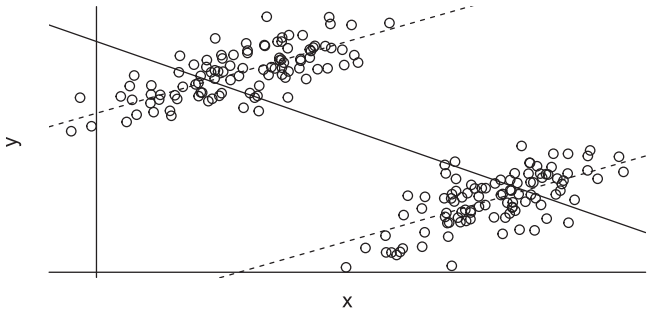

FIGURE 1 An extreme example under which blinded model selection can go wrong. The correct model (parallel dashed lines) shows a positive slope between y and x and a large difference in intercepts in the two arms, but blinded model selection using all data results in a negative slope between y and x

The blinded alternative method described above can fail spectacularly in very unrealistic settings. For example, consider the analysis of covariance setting of a continuous endpoint and a single, continuous covariate x in addition to the treatment indicator z. Figure 1 depicts a scenario in which the x values are dramatically different in the treatment and control arms and there is a huge treatment effect (difference in intercepts between the two dashed lines). The usual unblinded analysis estimates the common slope of the relationship between x and y as +1.02, but the blinded analysis yields an estimated slope of −1.24 (solid line). The resulting treatment effect estimate is greatly underestimated. Examples like this are very unlikely because they require a large difference between x values in the two arms and a huge treatment effect. The problem can be avoided altogether using an automated, unblinded selection of model followed by a randomization test on the entire process. Specifically, one could follow these steps: (1) re-randomize with the same method that was used in the actual trial, (2) apply the automated model selection procedure that includes covariates and the treatment variable, (3) compute the covariate-adjusted treatment effect estimate, (4) repeat steps 1–3 thousands of times, and (5) see where the observed treatment effect estimate from the original randomization lies with respect to this

randomization distribution. Confidence intervals can also be computed using the method described by Wang and Rosenberger (2020).

It would be interesting to see if the above randomization-based methods that account for covariates perform similarly to the methods studied by Benkeser *et al.* Also of interest is how marginal approaches compare with conditional methods such as logistic regression and proportional hazards regression in terms of efficiency.

Benkeser *et al.* should be congratulated for showing that robust and under-utilized methods improve efficiency and reduce sample sizes, relative to unadjusted methods, for trials of infectious diseases such as COVID-19. One small correction is that the primary outcome of Beigel *et al.* (2020) was time to recovery, not time to death.

ORCID

Michael A. Proschan https://orcid.org/0000-0002-9161-3739

REFERENCES

Beigel, J.H., Tomashek, K.M., Dodd, L.E., Mehta, A.K., Zingman, B.S., Kalil, A.C. et al. (2020) Remdesivir for the treatment of COVID-19–final report. *New England Journal of Medicine*, 383, 1813–1826.

Benkeser, D., Diaz, I., Luedtke, A., Segal, J., Scharfstein, D. and Rosenblum, M. (2021) Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment for binary, ordinal, and time-to-event outcomes. *Biometrics*. https://doi.org/10.1111/biom.13377. Epub ahead of print. PMID: 32978962; PMCID: PMC7537316.

Dodd, L., Follmann, D., Wang, J., Koenig, F., Korn, L.L., Schoergenhofer, C. et al. (2020) Endpoints for randomized controlled trials for COVID-19 treatments. *Clinical Trials*, 17, 472–482.

Gail, M.H., Tan, W.Y. and Piantadosi, S. (1988) Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75, 57–64.

Simon, R. and Simon, N.R. (2011) Using randomization tests to preserve type 1 error with response-adaptive and covariate-adaptive randomization. *Statistics and Probability Letters*, 81, 767–772.

Wang, Y. and Rosenberger, W. (2020). Randomization-based interval estimation in randomized clinical trials. *Statistics in Medicine*, 39, 2843–2854.

How to cite this article: Proschan MA. Discussion on "Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment for binary, ordinal, and time-to-event outcomes". *Biometrics*. 2021;77:1482–1484. https://doi.org/10.1111/biom.13493