Leveraging historical data for covariate adjustment in the analysis of randomized clinical trials

Authors: Samuel Branders[1], Alvaro Pereira[1], Guillaume Bernard[1], Marie Ernst[2], and Adelin Albert[2,3]

Affiliation(s):
[1] Tools4Patient SA, 1435 Mont-Saint-Guibert, Belgium
[2] Biostatistics, Medico-economic information Department, University Hospital of Liège, 4000 Liège, Belgium
[3] Department of Public Health, University of Liège, 4000 Liège, Belgium

Introduction and Motivation

The amount of potential covariates collected in clinical trials is continuously growing. However, the sample size of phase I/II studies strongly limits their use in practice.

Focusing on the estimation of the treatment effect, we tried to answer a few questions starting with:

'How to adjust for many (too many) covariates?"

"What is the optimum number of covariates to include in a clinical trial?"

Generative model of the data

The response model writes

$$Y = \mu + \gamma Z + U \tag{1}$$

where Z is the treatment variable and $U \sim \mathcal{N}(0, \sigma_u^2)$ is the error term.

The random variable U may in turn be expressed as a linear function of the covariates X, namely

$$U = \beta^T X + \varepsilon \tag{2}$$

where $X = (X_1, \ldots, X_p)^T$ is a vector of p covariates and the error term ε is assumed to be normally distributed $\mathcal{N}(0, \sigma_\varepsilon^2)$, independently of X.

We also define ν_p as the proportion of the variance of U explained by the covariates:

$$\nu_p := 1 - \frac{\sigma_\varepsilon^2}{\sigma_u^2}, \tag{3}$$

Variance of the estimated treatment effect

Without any covariate ($p = 0$) :

$$\mathrm{Var}(\hat{\gamma}_0 | z) = \frac{\sigma_u^2}{\sum_{i=1}^n (z_i - \bar{z})^2}. \tag{4}$$

With p covariates :

$$\mathrm{Var}(\hat{\gamma}_p | z, \mathbf{X}) = \frac{\sigma_\varepsilon^2}{(1 - \hat{R}^2_{z:\mathbf{X}}) \sum_{i=1}^n (z_i - \bar{z})^2} \tag{5}$$

where $\hat{R}^2_{z:\mathbf{X}}$ is the estimated multiple coefficient of determination of the regression of Z on the p covariates X.

Benefits of including covariates

Conditional on z and $\mathbf{X}$, a gain in the statistical precision of the estimated treatment effect is obtained if

$$\frac{\mathrm{Var}(\hat{\gamma}_p)}{\mathrm{Var}(\hat{\gamma}_0)} = \frac{1}{(1 - \hat{R}^2_{z:\mathbf{X}})} \frac{\sigma_\varepsilon^2}{\sigma_u^2} < 1. \tag{6}$$

If we assume that the covariates have independent and identical normal distribution, the expectation of this ratio and the inequality become :

$$RE_p := E_{Z,X}\left[\frac{1}{(1 - \hat{R}^2_{z:\mathbf{X}})} \frac{\sigma_\varepsilon^2}{\sigma_u^2}\right] = \frac{(n-3)}{(n-p-3)}(1 - \nu_p) < 1 \tag{7}$$

As a consequence, the p covariates included in the model improve the statistical precision of the estimator of γ if

$$\nu_p > \frac{p}{n-3} \qquad \text{or} \qquad p < (n-3)\nu_p. \tag{8}$$

Optimal number of covariates

Using the previous results, the optimal number of covariates can determined with

$$\underset{p}{\mathrm{argmin}} \frac{(n-3)}{(n-p-3)}(1 - \nu_p) \tag{9}$$

where ν_p is the population coefficient of determination of the regression with respect to the p most important covariates.

In particular, ν_p can be estimated from historical data of study in the same indication.

Composite covariates

The main problem with the use of covariates is the associated loss in degrees of freedom.

Assuming that historical data exists, we can go further and estimate the vector of covariates weights, β, directly from historical data.

The previous model simplifies as

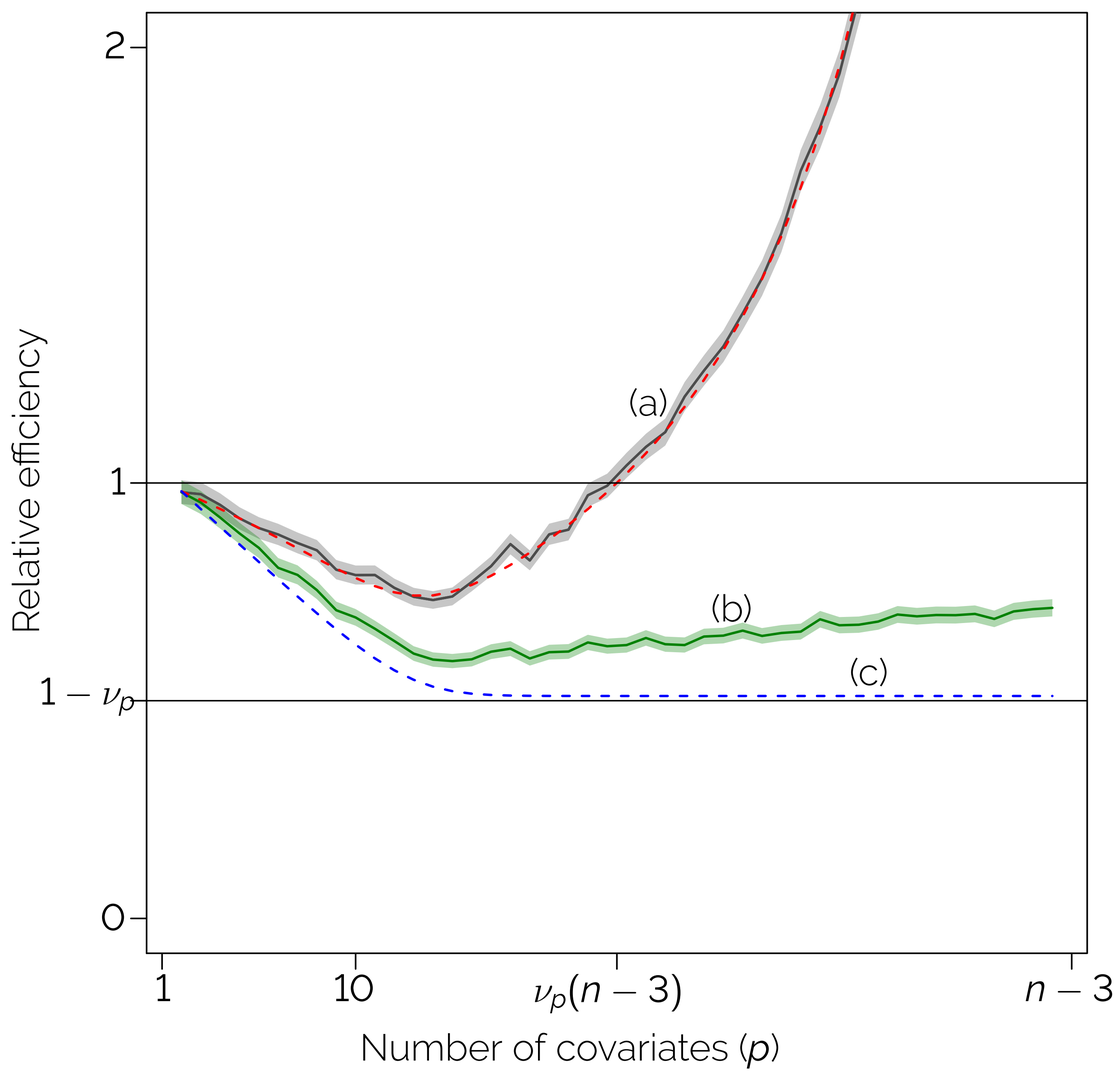
$$Y = \gamma Z + \beta W + \varepsilon. \tag{10}$$

where $W = f(X)$ as a composition of the p individual covariates, a composite covariate.

Simulation studies

Using simulations, we tested the composite covariate approach and its benefit.

Fig 1: Relative efficiency with respect to the estimation of the treatment effect without covariates.



(a) Mean relative efficiency of the p covariates and its expected value, $(n-3)(1-\nu_p)/(n-3-p)$.

(b) Mean relative efficiency of the composite covariate.

(c) Expected relative efficiency of an ideal composite covariate assuming $\nu_W = \nu_p$.

Conclusions

We showed how the variance of the estimated treatment effect evolves depending on the number of patients, covariates, and the variance explained by those covariates.

Then, using historical data, it's possible to estimate in advance the optimal number of covariates to maximize the expected precision of the treatment effect estimate.

We also showed how composite covariates could be used specifically to optimize the precision of the treatment effect estimation.

Using a composite covariate allows trading some explained variance to avoid the loss in degrees of freedom. The associated gain is particularly relevant when the sample size is small and the number of covariates is large.