Received: 5 November 2021 | Revised: 22 June 2022 | Accepted: 4 July 2022

DOI: 10.1002/bimj.202100349

Biometrical Journal

RESEARCH ARTICLE

On the relevance of prognostic information for clinical trials: A theoretical quantification

Sandra Siegfried[1] | Stephen Senn[2] | Torsten Hothorn[1]

[1]Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Zürich, Switzerland

[2]School of Health and Related Research, University of Sheffield, Sheffield, UK

Correspondence
Torsten Hothorn, Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Zürich, Switzerland.
Email: torsten.hothorn@r-project.org

Funding information
Horizon 2020 Research and Innovation Programme of the European Union, Grant/Award Number: 681094; Swiss State Secretariat for Education, Research and Innovation (SERI), Grant/Award Number: 15.0137; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 200021_184603

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

The question of how individual patient data from cohort studies or historical clinical trials can be leveraged for designing more powerful, or smaller yet equally powerful, clinical trials becomes increasingly important in the era of digitalization. Today, the traditional statistical analyses approaches may seem questionable to practitioners in light of ubiquitous historical prognostic information. Several methodological developments aim at incorporating historical information in the design and analysis of future clinical trials, most importantly Bayesian information borrowing, propensity score methods, stratification, and covariate adjustment. Adjusting the analysis with respect to a prognostic score, which was obtained from some model applied to historical data, received renewed interest from a machine learning perspective, and we study the potential of this approach for randomized clinical trials. In an idealized situation of a normal outcome in a two-arm trial with 1:1 allocation, we derive a simple sample size reduction formula as a function of two criteria characterizing the prognostic score: (1) the coefficient of determination R^2 on historical data and (2) the correlation ρ between the estimated and the true unknown prognostic scores. While maintaining the same power, the original total sample size n planned for the unadjusted analysis reduces to $(1 - R^2\rho^2) \times n$ in an adjusted analysis. Robustness in less ideal situations was assessed empirically. We conclude that there is potential for substantially more powerful or smaller trials, but only when prognostic scores can be accurately estimated.

KEYWORDS
clinical trials, covariate adjustment, machine learning, prognostic covariates, sample size reduction

1 | INTRODUCTION

Randomized controlled trials (RCTs) are the gold standard design for the estimation of an average treatment effect of some novel intervention. The high level of evidence deducible from such a study, however, comes at a high price: Large sample sizes are often required to demonstrate an anticipated treatment effect with sufficient power. This not only renders many RCTs financially intensive, but also raises ethical considerations. An important goal of methodological research is therefore the development of methods allowing for a substantial reduction of the overall sample size or to estimate the treatment effect with higher precision from equally large trials.

In many contexts, individual patient data from large cohort studies or previously conducted RCTs have been collected with great effort over long periods of time. Such data contain valuable information about the course of a disease under standard of care or even in untreated patient populations. When planning a novel RCT, the questions "if" and "how" such prognostic information can be leveraged to increase precision or to reduce the necessary future sample size arise naturally.

Many contributions to contemporary RCT methodology can be understood as attempts to solve this common problem. Information borrowing, propensity score matching and adjustment, stratification, and covariate adjustment are the main strands of research concentrating on the "how" part of the question. We focus on the "if" aspect and try to identify conditions allowing trials to be smaller through incorporation of historical prognostic information. In an idealized normal model, we derive a simple relationship between the strength of prognostic information contained in historical controls, the quality of a prognostic score capturing this information, and the reduction in total sample size or gain in precision achievable by adjusting for such a prognostic score in an RCT.

A prognostic score represents a baseline risk in terms of a summary score of observed covariates (Breslow, 1979). More specifically, the score quantifies the expected response under control conditions, estimated from reference data, for example, historical control data. The concept of prognostic scores can thus be utilized to collapse large number of covariates, and potentially high-dimensional or unstructured information, in a composite score. In clinical practice, prognostic scores aim to provide a tool for risk stratification, for example, for clinical behavior of a disease (e.g., in prostate cancer, Kreuz et al., 2020) or in the intensive care unit (e.g., FOUR score, Wijdicks et al., 2005). Prognostic scores have been used for defining strata in clinical trials (Cellini et al., 2019; Herrera et al., 2020), constituting a simple form of incorporating historical information in the trial design. In observational studies, Hansen (2008) formalized the concept of adjusting with respect to a prognostic score, which has since been studied for various applications (Arbogast & Ray, 2009, 2011; Hajage et al., 2017; Pfeiffer & Riedl, 2015; Wyss et al., 2016). Although the idea of adjusting with respect to historical prognostic scores in clinical trials was briefly sketched in an earlier paper by Cox (1982) and has been suggested for heterogeneous treatment effect estimation (Kent et al., 2020), an in-depth development of this principle has been published only recently (Anonymous, 2022; Branders et al., 2021; Schuler et al., 2021, which was under consultation for qualification opinion by the European Medicines Agency [EMA]).

Prognostic score methods have strong ties to stratification and covariate adjustment, where, in practice, little is known about the actual extent of the efficiency gained by stratification (Kernan et al., 1999) or covariate adjustment (Robinson & Jewell, 1991; Steingrimsson et al., 2017). Similar to information borrowing or propensity score matching and adjustment, the prognostic score dynamically leverages historical information.

In our work, we explore this idea in an exemplary setup to quantify the benefits, "if" prognostic information is leveraged in the statistical analysis. We present a simple and general situation in Section 2, and contrast conditions determining the potential benefits when employing this approach in Section 3.

2 | METHODS

We consider a simple two-arm RCT aiming to estimate the effect of a treatment on some continuous primary outcome $Y \in \mathbb{R}$. In the trial, patients were randomly assigned to either the treatment, $z = 1$, or the control arm, $z = 0$. For each patient a set of patient characteristics $\boldsymbol{X} \in \chi$ were retrieved at baseline, from potentially high-dimensional, structured, or unstructured information. The prognostic score is defined in terms of an unknown function $s : \chi \rightarrow \mathbb{R}$ collapsing the k baseline covariates in $\boldsymbol{X}$. Assuming the outcome Y stems from a normal distribution, we study the following data-generating process (DGP)

$$Y = \alpha + \beta z + \{\pi s(\boldsymbol{X}) + \sqrt{\sigma^2 - \pi^2}\varepsilon\} \sim \mathrm{N}(\alpha + \beta z, \sigma^2), \tag{1}$$

where α is the intercept parameter and β the treatment effect we wish to estimate. The unexplained variability σ^2 is decomposed into a structured error term,

$$\{\pi s(\mathbf{X}) + \sqrt{\sigma^2 - \pi^2}\varepsilon\} \sim \mathrm{N}(0, \sigma^2), \mathbf{X} \perp\!\!\!\perp \varepsilon, \tag{2}$$

consisting of a mixture distribution of a prognostic score $s(\boldsymbol{X}) \sim \mathrm{N}(0, 1)$, which follows a standard normal distribution by assumption, and an independent standard normal residual $\varepsilon \sim \mathrm{N}(0, 1)$. The parameter $\pi \in [0, \sigma]$ governs the fraction of variability explained by the prognostic score $s(\boldsymbol{X})$.

The standard deviation of the residual, $\sqrt{\sigma^2 - \pi^2}$, depends on π, such that the variance σ^2 of the structured error term (2) is constant. For $\pi = 0$, the residual variance is σ^2 and the prognostic score does not impact the outcome in any way. For $\pi = \sigma$, the prognostic score $s(\boldsymbol{X})$ accounts for the total variability and the residual variance is zero. Values of $\pi \in (0, \sigma)$ indicate DGPs with different signal-to-noise ratios regarding the prognostic score $s(\boldsymbol{X})$. Large values of $\pi s(\boldsymbol{X})$ are associated with large values of the outcome Y, in both the treatment and control groups.

In rare cases, the prognostic score function $\pi s()$ might be known and $\pi s(\boldsymbol{x})$ can be used as an offset in (1), when $\boldsymbol{X} = \boldsymbol{x}$ was observed for patients in the trial. The standard error of the treatment parameter estimate, $\hat{\beta}$, and thus also the sample size necessary to demonstrate a certain clinically relevant effect, only depend on the residual variance $\sigma^2 - \pi^2$ in this case. Typically, neither π nor the prognostic score $s(\boldsymbol{x})$ are available and need to be estimated. Sometimes, it is appropriate to assume a linear model $\pi s(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\gamma}$, where an adjusted estimate for the treatment effect β is computed from simultaneous estimation with $\boldsymbol{\gamma}$. Using trial data, the joint estimation of the treatment parameter β and $\pi s(\boldsymbol{x})$ is much more difficult, inefficient, or even impossible for high-dimensional (e.g., high-throughput, biomarker data) or unstructured (e.g., clinical notes and reports) covariates $\boldsymbol{X}$ (Zhang & Ma, 2019), thus potentially necessitating an independent sample for the estimation of $\pi s(\boldsymbol{x})$.

We are interested in the setup, where one was able to obtain an estimate, $\mathfrak{s}(\boldsymbol{x}) = \widehat{\pi s}(\boldsymbol{x})$, of $\pi s(\boldsymbol{x})$ either from the literature or from historical control data. The latter situation received some interest recently (Branders et al., 2021; Glynn et al., 2012; Schuler et al., 2021; Wyss et al., 2014). Assuming one has access to data from past trials on the same outcome Y and covariates $\boldsymbol{X}$ for control patients, $z = 0$, many statistical and machine learning procedures, for example, random forests and neural networks, can be used to estimate the prognostic score function from the conditional mean $\mathfrak{s}(\boldsymbol{x}) = \widehat{\pi s}(\boldsymbol{x}) = \hat{\mathbb{E}}(Y \mid \boldsymbol{X} = \boldsymbol{x}, z = 0) - \hat{\alpha}$ (with some estimate $\hat{\alpha}$ of the model's intercept parameter). Model (1) for historical controls ($z = 0$) regressing on $\boldsymbol{X} = \boldsymbol{x}$ is associated with an ideal percent explained variability of $R^2 = {}^{\pi^2}/_{\sigma^2}$, or, more formally

$$R^2(g) := 1 - \frac{\mathbb{V}\{Y - g(\mathbf{X}) \mid z = 0\}}{\mathbb{V}\{Y \mid z = 0\}} \quad \text{for some regression function } g \text{ and}$$

$$R^2 = R^2(\pi s) = 1 - \frac{\sigma^2 - \pi^2}{\sigma^2} = \frac{\pi^2}{\sigma^2}.$$

Instead of studying properties of specific estimators, we make an assumption about the joint distribution of the estimated and the true prognostic scores in terms of a correlation coefficient $\rho \in [0, 1]$ for the relevant situation $\pi > 0$,

$$(\mathfrak{s}(\boldsymbol{X}), \pi s(\boldsymbol{X})) \sim \mathrm{N}_2\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \pi^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]. \tag{3}$$

The setup $\rho = 0$ corresponds to a failed attempt to estimate the prognostic score on historical data. For $\rho = 1$, we obtained an oracle $\mathfrak{s}(\boldsymbol{X}) = \pi s(\boldsymbol{X})$, possibly from some very big database or from expert elicitation. More realistically, values $\rho \in (0, 1)$ describe how well the prognostic model $\mathfrak{s}(\boldsymbol{X})$ characterizes the prognostic score $\pi s(\boldsymbol{X})$; the corresponding mean-squared error is

$$\mathbb{E}[\{\pi s(\boldsymbol{X}) - \mathfrak{s}(\boldsymbol{X})\}^2] = 2\pi^2(1 - \rho).$$

For the sake of completeness, we introduce a symbol for the out-of-sample (OOS) percent explained variability one would obtain, for example, by cross-validation or an additional test sample evaluating the prognostic score $\mathfrak{s}$ fitted to historical data only in model (1) :

$$\begin{aligned} R^2_{\text{OOS}} = R^2(\hat{s}) &= 1 - \frac{\mathbb{V}\{Y - \hat{s}(\boldsymbol{X}) \mid z = 0\}}{\mathbb{V}\{Y \mid z = 0\}} \\ &= 1 - \frac{\mathbb{E}[\{\pi s(\boldsymbol{X}) - \hat{s}(\boldsymbol{X})\}^2] + \sigma^2 - \pi^2}{\sigma^2} \\ &= (2\rho - 1) \times R^2. \end{aligned}$$

The predicted variance reduction for the trial, following Borm et al. (2007) and also more recently Branders et al. (2021) and Schuler et al. (2021), would then be $1 - \widehat{R^2}_{\text{OOS}}$.

In our simple setup, it is straightforward to see that one can replace the unknown prognostic score $\pi s(\boldsymbol{X})$ by $\rho\hat{s}(\boldsymbol{X})$ in (1) without changing the distribution of the outcome,

$$\begin{aligned} Y &= \alpha + \beta z + \pi s(\boldsymbol{X}) + \sqrt{\sigma^2 - \pi^2}\varepsilon \sim \text{N}(\alpha + \beta z, \sigma^2) \\ &\overset{d}{=} \alpha + \beta z + \rho\hat{s}(\boldsymbol{X}) + \sqrt{\sigma^2 - \pi^2\rho^2}\varepsilon \sim \text{N}(\alpha + \beta z, \sigma^2). \end{aligned}$$

For the trial patients, this change means that treating $\hat{s}(\boldsymbol{X}) \in \mathbb{R}$ as a single observable and random covariate with unknown regression coefficient ρ leads to a reduction of the residual variance from σ^2 (in a model $Y \mid z \sim \text{N}(\alpha + \beta z, \sigma^2)$ ignoring prognostic information) to $\sigma^2 - \pi^2\rho^2$ (in a model $Y \mid z, \hat{s}(\boldsymbol{x}) \sim \text{N}(\alpha + \beta z + \rho\hat{s}(\boldsymbol{x}), \sigma^2 - \pi^2\rho^2)$ adjusting for prognostic information $\hat{s}(\boldsymbol{x})$) whenever $\pi > 0$ and $\rho > 0$. At the price of estimating one additional parameter ρ in the linear model $Y \mid z, \hat{s}(\boldsymbol{x}) \sim \text{N}(\alpha + \beta z + \rho\hat{s}(\boldsymbol{x}), \sigma^2 - \pi^2\rho^2)$, one can expect a considerable reduction of the residual variance, and therefore more powerful tests and confidence intervals for β, when employing this method of adjustment. The fraction

$$\frac{\text{residual variance } Y \mid z, \hat{s}(\mathbf{x})}{\text{residual variance } Y \mid z} = \frac{\sigma^2 - \pi^2\rho^2}{\sigma^2} = 1 - R^2\rho^2 \tag{4}$$

of the residual variances with and without adjustment for prognostic information approximately corresponds to the fraction of necessary sample sizes to demonstrate a specific clinically relevant treatment effect. This holds for any nominal significance level and power because the sample size of the t-test decreases linearly with residual variance when the estimation of the additional parameter ρ is ignored. This issue and the connection to ANCOVA is discussed in Section 4. Equivalently, for fixed sample size n the precision of the treatment effect estimate increases as the residual variance decreases. It should be noted that the classical "design factor" $1 - \widehat{R^2}_{\text{OOS}}$ (Borm et al., 2007; Branders et al., 2021; Schuler et al., 2021) is biased in our setup, because $1 - R^2_{\text{OOS}} = 1 - (2\rho - 1)R^2 \neq 1 - \rho^2 R^2$. This discrepancy will be demonstrated empirically in Section 3.

The setup also captures a potential distribution drift from the historical to the trial data: Even if $\hat{s}(\boldsymbol{X})$ is a very precise estimator of the true prognostic score on the historical data, a considerable lack of fit on the trial data, and thus a small ρ, might be due to a temporal drift in the prognostic score, which applies to trial but not historical patients.

In contrast to classical covariate adjustment, the relationship between $\boldsymbol{X}$ and Y can be highly nonlinear or unstructured in our studied setup, for example, when $\boldsymbol{X}$ represents image data and a complex deep neural network is used to obtain $\hat{s}(\boldsymbol{X})$. Still only a single additional parameter ρ has to be estimated in addition to the treatment effect β from the present trial data. The type I error for hypothesis tests on β is maintained, assuming the test procedure deals with random covariates in an appropriate way, and thus lack of type I error control reported for Bayesian borrowing procedures (Kopp-Schneider et al., 2020) is avoided here.

The most important question is: When does it actually theoretically pay off to leverage prognostic information by incorporating prognostic scores $\hat{s}(\boldsymbol{x})$ estimated on historical data? We assess this question theoretically and empirically for specific values $R^2 = {}^{\pi^2}\!/_{\sigma^2} \in (0, 1)$ and $\rho \in (0, 1)$ in Section 3.1. Furthermore, we study the impact of deviations from the rather strict distributional assumption (3) on the prognostic score and its estimate in Section 3.2. Unfortunately, it is impossible to estimate the theoretical model parameters ρ and R^2 from historical or contemporary data. This makes it difficult to provide practical rules of thumb guiding an informed decision on leveraging prognostic scores in the analysis of clinical trial data.

3 | RESULTS

3.1 | Illustration of theoretical result

The fraction (4) of residual variances with and without adjustment $1 - R^2\rho^2$ for values of $R^2 \in (0, 1)$ and $\rho \in (0, 1)$ lead to the following interpretation: For a clinical trial powered for the demonstration of a certain clinically relevant effect β in a normal model with variance σ^2 with a specific nominal level and power, the planned sample size n can be reduced to $(1 - R^2\rho^2) \times n$ through adjustment for prognostic information. For example, with $R^2 = 0.5$ on a large historical data set resulting in a very precise estimate $\mathfrak{s}(\boldsymbol{X})$ of $\pi s(\boldsymbol{X})$ with $\rho = 0.8$, say, only $(1 - 0.5 \times 0.8^2) \times 100\% = 68\%$ of the original sample size n would be required in an adjusted analysis. Substantial reductions by more than 20% of the original sample size (i.e., $1 - R^2\rho^2 < 0.8$) can only be expected for $R^2 > 0.3$ and rather large values of ρ. The higher R^2, the less precision of the estimate $\mathfrak{s}(\boldsymbol{X})$ is necessary to achieve the same level of reduction. For situations with either small R^2 on the historical data and/or small historical sample sizes $\mathfrak{n}$ resulting in smaller values of ρ, expected sample size reductions of less than 10% (i.e., $1 - R^2\rho^2 > 0.9$) suggest that accounting for prognostic information might not be worth the effort.

3.2 | Sensitivity analysis

To study the impact of deviations from the distributional assumption (3), we contrasted the above presented results with a more complex DGP. For the prognostic score, we employed the process

$$s(\boldsymbol{X}) = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon, \tag{5}$$

arising from Friedman' regression equation 1 (Friedman, 1991) with $\boldsymbol{X} \sim \mathrm{U}(0, 1)^{10}$ and $\varepsilon \sim \mathrm{N}(0, 1)$.

We simulated historical control data ($z = 0$) of varying sample size $\mathfrak{n} = 50$, 100, and 10,000 as well as trial data with sample size $n = 1000$ from DGP (1) with $\sigma^2 = 1$ for different values of $\pi \in (0, 1)$ and repeated the experiment 1000 times. We estimated the prognostic model $\mathfrak{s}(\boldsymbol{X})$ from the simulated historical control data with a random forest and fitted a normal linear regression model for the treatment effect to the trial data and a model additionally adjusting for the prognostic score estimate $\mathfrak{s}(\boldsymbol{x})$. The additive DGP (5) cannot be exactly recovered by a random forest model and thus an additional bias is introduced. The precision of the random forest $\hat{\rho}$, which was estimated from data, increases with larger values of R^2 and sample size $\mathfrak{n}$ (Figure 2). The results in Figure 1 convey similar findings as obtained theoretically. The residual variance when adjusting for the prognostic score estimated from historical data decreases with higher R^2, which translates into higher precision of the treatment effect estimates $\hat{\beta}$ (Figure 3).

3.3 | Comparison of predicted and empirical variance reduction

We further compared the variance reduction achieved by prognostic score adjustment as predicted by the "design factor" $1 - \widehat{R^2}_{\mathrm{OOS}}$ (Borm et al., 2007), using the estimated $\widehat{R^2}_{\mathrm{OOS}}$ from the prognostic random forest model on historical data, to the empirical variance reduction $1 - \rho^2 R^2$ in our setup. For the DGP in Section 3.2, random forests' $\widehat{R^2}_{\mathrm{OOS}}$ was estimated using a large evaluation data set (OOS). The true R^2 was calculated using π^2/σ^2.

The lines in Figure 1 contrast the variance reduction predicted by the "design factor" $1 - \widehat{R^2}_{\mathrm{OOS}}$ (Borm et al., 2007) and $1 - \rho^2 R^2$ (Fraction 4) with the variance reduction achieved empirically (boxplots). The latter variance reduction fits the empirical results very closely, whereas the "design factor" is biased and underestimates the actual observed variance reduction. For very large historical sample sizes, the variance reduction is well described by the "design factor."

3.4 | Illustration

A recent study by Goemans et al. (2020) reported on the development of a prognostic score for timed four-stair climb in Duchenne muscular dystrophy patients and discussed its potential benefits in terms of design and analysis of future trials. The explained variability (R^2_{OOS}) in the prognostic model was described to be maximally 36%, which according to

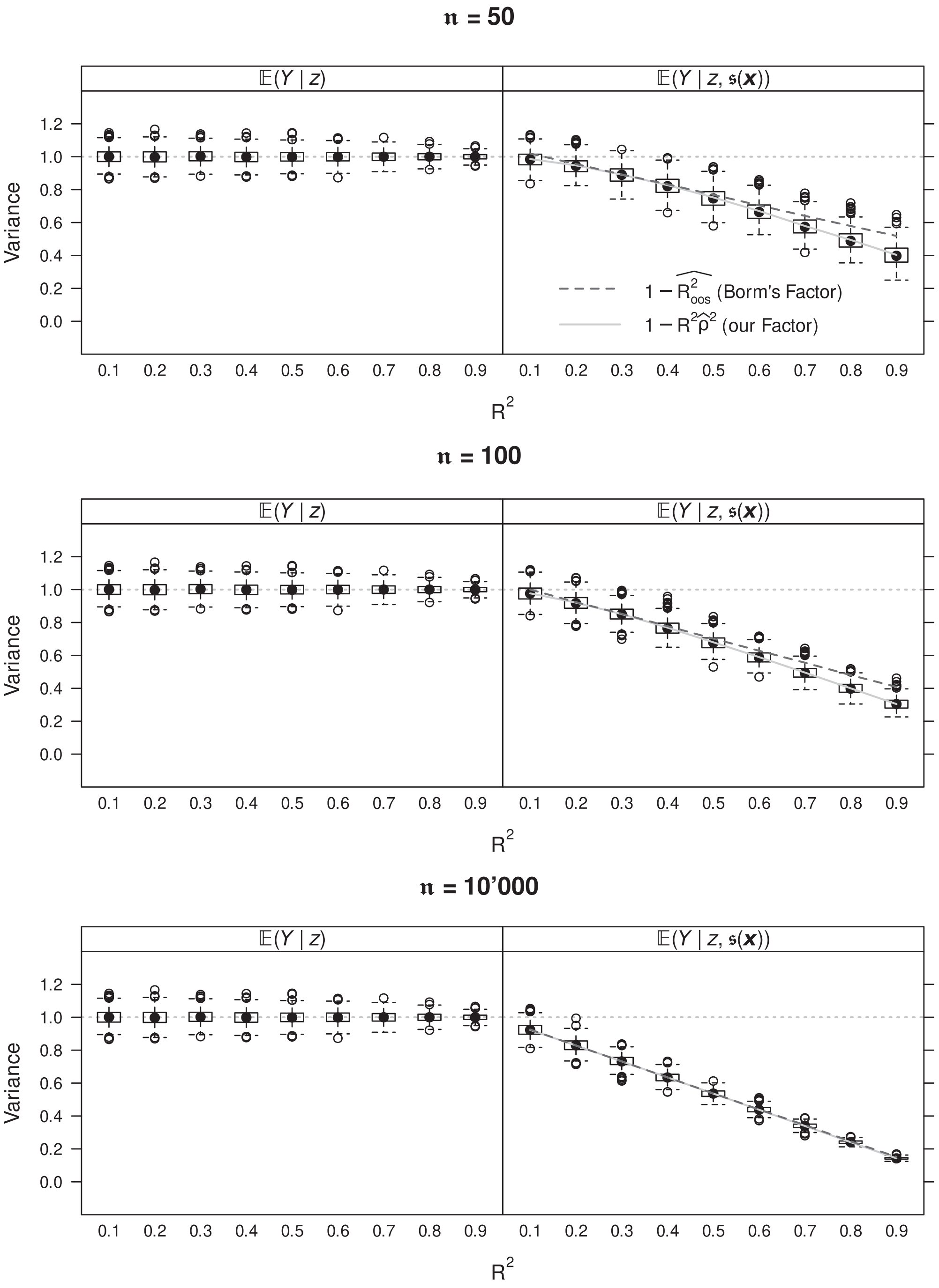

FIGURE 1 Simulated fraction of residual variances in a model with prognostic score as defined in (5). The fractions are shown for the normal linear model regressing on the treatment effect ($\mathbb{E}(Y \mid z)$; left), and the model additionally adjusting for the prognostic score estimate ($\mathbb{E}(Y \mid z, \hat{\mathfrak{s}}(\boldsymbol{x}))$; right) for various values of $R^2 = \pi^2/\sigma^2$ and different sample sizes $\mathfrak{n} = 50, 100$, and 10,000. The light gray line depicts the theoretical fraction $1 - R^2\hat{\rho}^2$, with the precision of the random forest $\hat{\rho}$ estimated from the data. The variance reduction predicted by the "design factor" (Borm et al., 2007) is shown as dashed dark gray line

the "design factor" would allow for a variance reduction to 64% of the unadjusted analysis when employing prognostic score adjustment. The empirical reduction however is difficult to quantify, because in practice R^2 and ρ are unknown. Based on our derivations, the reported reduction to 64% would only be attainable in absence of distributional drift, for example, with $\rho = 0.9$ and corresponding $R^2 = 0.44$. However, for trial data deviating from the historical training data and thus smaller values of ρ, this variance reduction would require larger values of R^2, for example, with $\rho = 0.7$ the corresponding $R^2 = 0.73$ would be needed.

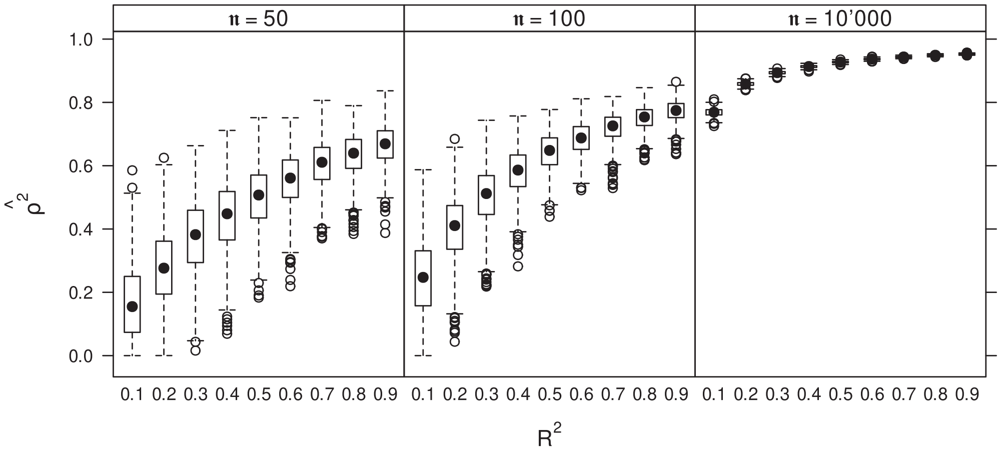

FIGURE 2 Characteristics of the employed prognostic model $\hat{\mathfrak{s}}(\boldsymbol{X})$. The precision of the random forest $\hat{\rho}$ estimated from the data are shown along different values of $R^2 = \pi^2/\sigma^2$ for different historical sample sizes $\mathfrak{n} = 50, 100$, and 10,000

4 | RELATION TO COVARIATE ADJUSTMENT

In the analysis of covariance (ANCOVA) framework, an interesting practical question is when it will be more beneficial to directly adjust for prognostic variables instead of adjusting for a prognostic score, or even not to adjust at all (Lesaffre & Senn, 2003). We shall discuss this issue in more detail.

Suppose we have n subjects in total and $k \geq 2$ prognostic covariates. (The lower bound is set at 2 since the case $k = 1$ is without interest.) The loss due to nonorthogonality, which we refer to as the *imbalance effect* is a random variable depending on the observed imbalance in the trial. However, choosing whether to fit the score or the covariates based on an inspection of the data has the danger of increasing the type 1 error rate. Thus, there is merit in making a prespecified choice of a model, which, in any case, is in line with ICHe9 recommendations. It can be shown, however, that the expected imbalance effect due to fitting k covariates compared to 1 is $(n-4)/(n-3-k)$. On the other hand, the expected inflation in the mean square error (MSE), which we refer to as the *MSE effect*, due to fitting a score based on historical data rather than the k covariates on which it is based is $\sigma_1^2/\sigma_k^2 \geq 1$, where the numerator is the expected MSE for the prognostic score and the denominator the corresponding MSE with all covariates fitted. Thus, by comparing the MSE effect to the expected imbalance effect, one can make a decision. Note that a third element to consider is that the residual degrees of freedom for error will lead to the t-table having to be entered at a less favorable point, the more covariates are fitted. As is discussed in the Appendix this further effect, which we refer to as *second-order precision*, will favor the prognostic score.

In summary, when the trial sample size n is large and only a few prognostic variables are studied, using ANCOVA without any involvement of historical data should be preferred (Borm et al., 2007; Cox & McCullagh, 1982; Pocock et al., 2002), whenever the linearity assumption is justified. In situations where either the trial sample size n is relatively small, many and potentially unstructured prognostic variables shall be adjusted for, and a large set of $\mathfrak{n}$ historical patient records is available, it seems preferable to adjust for the prognostic score in situations where $R^2 > 0.3$, because only one additional parameter needs to be estimated in a classical statistical model.

5 | DISCUSSION

In our work, we studied the question, in what situations leveraging prognostic information actually pays off in practice. We presented a simple and general setup in Section 2, allowing us to assess the theoretical properties of this adjustment method without making strong distributional assumptions or limiting it to specific estimators.

In Section 3.1, we quantified the maximally attainable benefit when adjusting for a prognostic score analytically, and contrasted our findings with a more complex set-up in Section 3.2. The results suggest that leveraging prognostic baseline covariates reduces residual variability, however the magnitude of this reduction might often be irrelevant in practice. These situations can be characterized by small historical samples sizes (and as a result smaller ρ) and/or small R^2 of the prognostic model on historical data.

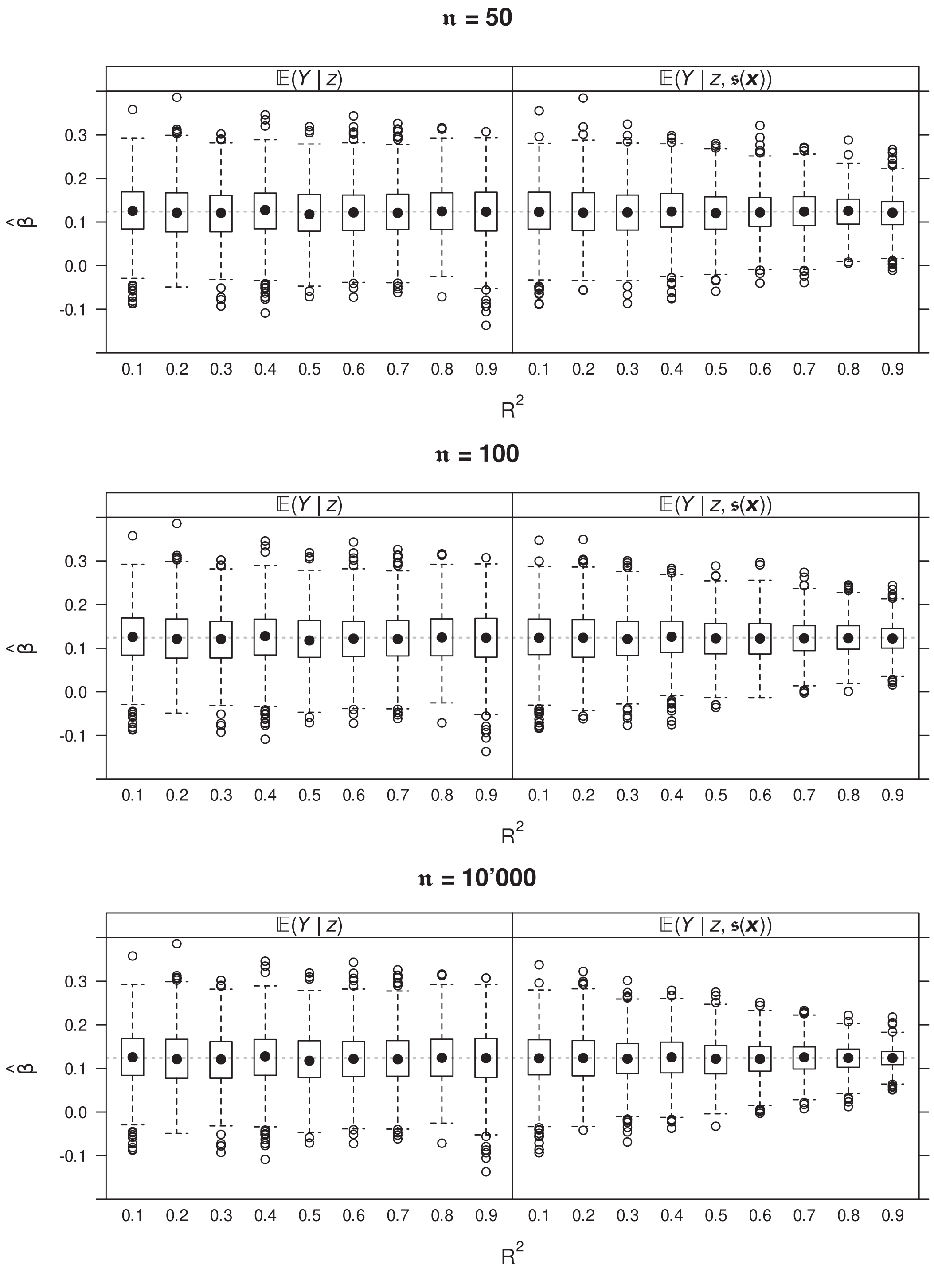

FIGURE 3 Simulated distribution of the treatment effect estimate. The treatment effect estimates $\hat{\beta}$ from the normal linear model regressing on the treatment effect ($\mathbb{E}(Y \mid z)$; left), and the model additionally adjusting for the prognostic score estimate ($\mathbb{E}(Y \mid z, \hat{\mathfrak{s}}(\boldsymbol{x}))$; right) are shown for various values of $R^2 = {}^{\pi^2}/_{\sigma^2}$ and different sample sizes $\mathfrak{n} = 50, 100$, and 10,000. The true treatment effect $\beta = 0.12$ is indicated by the horizontal line

As a rough rule of thumb, sample size reductions of more than 20% are achievable with an $R^2 > 0.3$ on historical controls when there is a high confidence in the prognostic score, with $\rho > 0.8$ say, requiring a large number of historical controls and the absence of drift in $s(\boldsymbol{X})$. When there is more uncertainty regarding the prognostic score, with $\rho \approx 0.6$ for example, an $R^2 > 0.5$ is necessary to obtain a 20% reduction in total sample size. Likewise, the corresponding increase in precision of the treatment effect estimate can be considered for fixed samples sizes. It depends on the context whether or not such an increase is relevant: It might be a game-changer in one setup but only marginally interesting in other situations. Athey and Imbens (2017) point out that randomization is sufficient for lack of bias in $\hat{\beta}$ from including or excluding a prognostic score, however "the gain in precision is often modest."

While it is easy to estimate R^2_{OOS} for historical controls, estimating our model parameters R^2 and ρ is less straightforward. One possibility would be to perform an interim analysis regressing the outcome Y on the prognostic score $\mathfrak{s}(\boldsymbol{X})$ on the trial controls $(Y, \boldsymbol{X}, z = 0)$, which, after appropriate standardization such that $\mathbb{V}(\mathfrak{s}(\boldsymbol{X})) = 1$, gives an estimate $\widehat{\pi\rho}$ for $\pi\rho$, which, together with an estimate of the residual variance σ^2, can be plugged into (4). In the absence of information about ρ, our interpretation of the theoretical results presented here is that trial designers should definitively look into the possibility of adjusting for an established prognostic score when its R^2_{OOS} has been demonstrated to exceed 0.5.

These findings are in agreement with earlier results quantifying the impact of covariate adjustment on the necessary sample size in clinical trials. Adjusting for a single numeric covariate X_1 is a special case of our model with $\pi s(\boldsymbol{X}) = \pi X_1$ and $\rho \equiv 1$, resulting in a "design factor" of $1 - R^2$, meaning a sample size reduction to $(1 - R^2) \times 100\%$ of original sample size is possible (Borm et al., 2007; Cox & McCullagh, 1982; Pocock et al., 2002). This "design factor" however disregards that the covariate (or equivalently the prognostic score) might be measured with error ρ or that there might be potential distribution drift.

Although accounting for prognostic information through adjustment for $\mathfrak{s}(\boldsymbol{X})$ seems rather unorthodox, a simpler version known as poststratification is well established. For two strata, the prognostic score $\mathfrak{s}(\boldsymbol{X}) \in \{0, 1\}$ is an indicator for the patient's stratum, ρ an unknown prognostic parameter, typically estimated from trial data. The rational is the same: leveraging information from historical controls (used to define reasonable strata) for reducing the residual variance while safeguarding against distribution shift or incorrectly specified strata. If available, such information further can be employed to randomize patients into more homogeneous subgroups.

6 | FUTURE RESEARCH AND CONCLUSION

If and how historical controls can be leveraged for future clinical trials has been discussed extensively, yet the debate did not converge to some consensus. With the increasing application of machine learning in clinical research, the general hope is that these techniques will eventually help to design more efficient trials. Prognostic score adjustment seems to play an important role toward this goal (Branders et al., 2021; Kent et al., 2020; Schuler et al., 2021). New sample size planning instruments are required, which properly take into account a realistic assessment of the prognostic value of existing or new scores for trial patients. Our contribution can only serve as a best-case benchmark scenario future methodological developments can be compared to.

An extension to nonnormal models is not straightforward. From a computational point of view, the estimation of prognostic scores on appropriate scales (log-odds or log-hazard ratios, for example) is possible by application of some machine learning procedures, for example, in model-based boosting (Bühlmann & Hothorn, 2007; Ridgeway, 1999; Schmid et al., 2011). Adjusting for such prognostic scores in logistic, proportional odds, or proportional hazards regression models will lead to increasing power for testing the null hypothesis $\beta = 0$ at the price of changing the interpretation of the treatment effect estimate $\hat{\beta}$ from a marginal to a conditional one (Daniel et al., 2021; Ford et al., 1995; Ford & Norrie, 2002; Hernández et al., 2004; Robinson & Jewell, 1991), owing to the fact that, unlike in nonlinear models, $\pi s(\boldsymbol{X})$ can be absorbed into the error term (2) in the linear model (1).

In summary, the lack of ability to estimate the relevant parameters R^2 and ρ on historical data, the inability to a priori assess potential drift between historical and future controls, the associated uncertainties when planning the sample size for a future trial, and the low cost of prognostic score adjustment in the final analysis suggest a pragmatic approach. Instead of a priori factoring in a certain sample size reduction, one should treat the potential for power increase by means of prognostic score adjustment as a nest egg. For traditionally planned trials aiming at differences in means, post hoc prognostic score adjustment with an insufficient score neither affects size nor power of the final analysis. However, post hoc adjustment with an informative prognostic score holds some potential for power increase without making strong a priori commitments.

ACKNOWLEDGMENTS

This work was supported by Swiss National Science Foundation, grant 200021_184603. Torsten Hothorn additionally acknowledges funding from the Horizon 2020 Research and Innovation Programme of the European Union under grant agreement number 681094, and is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0137. The authors thank Maria-Eleni Syleouni for initial simulation experiments testing the power of prognostic score adjustment in her master's thesis.

Open access funding provided by Universitat Zurich.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable – no new data generated.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

AUTHOR CONTRIBUTIONS

Sandra Siegfried drafted the paper, contributed to the theoretical part, and performed empirical experiments. Stephen Senn identified relevant earlier contributions and contributed the connection to ANCOVA provided in the Appendix. Torsten Hothorn designed the study and developed the model. All authors revised and approved the final version.

ORCID

Sandra Siegfried https://orcid.org/0000-0002-7312-1001
Torsten Hothorn https://orcid.org/0000-0001-8301-0471

REFERENCES

Anonymous (2022). *Draft qualification opinion for prognostic covariate adjustment (PROCOVA$^{\text{TM}}$)*. EMA/SA/0000059571, European Medicines Agency. https://www.ema.europa.eu

Arbogast, P. G., & Ray, W. A. (2009). Use of disease risk scores in pharmacoepidemiologic studies. *Statistical Methods in Medical Research*, *18*(1), 67–80. https://doi.org/10.1177/0962280208092347

Arbogast, P. G., & Ray, W. A. (2011). Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. *American Journal of Epidemiology*, *174*(5), 613–620. https://doi.org/10.1093/aje/kwr143

Athey, S., & Imbens, G. (2017). The econometrics of randomized experiments. In A. V. Banerjee, and E. Duflo, (Eds.), *Handbook of economic field experiments*, (Vol. 1, pp. 73–140). North-Holland. https://doi.org/10.1016/bs.hefe.2016.10.003

Borm, G. F., Fransen, J., & Lemmens, W. A. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of Clinical Epidemiology*, *60*(12), 1234–1238. https://doi.org/10.1016/j.jclinepi.2007.02.006

Branders, S., Pereira, A., Bernard, G., Ernst, M., Dananberg, J., & Albert, A. (2022). Leveraging historical data to optimize the number of covariates and their explained variance in the analysis of randomized clinical trials. *Statistical Methods in Medical Research*, *31*(2), 240–252. https://doi.org/10.1177/09622802211065246

Breslow, N. (1979). Statistical methods for censored survival data. *Environmental Health Perspectives*, *32*, 181–192. https://doi.org/10.1289/ehp.7932181

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*(4), 477–505. https://doi.org/10.1214/07-sts242

Cellini, F., Manfrida, S., Deodato, F., Cilla, S., Maranzano, E., Pergolizzi, S., Arcidiacono, F., Di Franco, R., Pastore, F., Muto, M., Borzillo, V., Donati, C. M., Siepe, G., Parisi, S., Salatino, A., D'Agostino, A., Montesi, G., Santacaterina, A., Fusco, V., … Corvò, R. (2019). Pain REduction with bone metastases STereotactic radiotherapy (PREST): A phase III randomized multicentric trial. *Trials*, *20*(1), 1–7. https://doi.org/10.1186/s13063-019-3676-x

Cox, D. R. (1982). Randomization and concomitant variables in the design of experiments. In G. Kallianpur, P. Krishnaiah, and J. Ghosh, (Eds.), *Statistics and probability: Essays in honor of CR Rao*, (pp. 197–202). North-Holland.

Cox, D. R., & McCullagh, P. (1982). A biometrics invited paper with discussion. some aspects of analysis of covariance. *Biometrics*, *38*(3), 541–561. https://doi.org/10.2307/2530040

Daniel, R., Zhang, J., & Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, *63*(3), 528–557. https://doi.org/10.1002/bimj.201900297

Ford, I., & Norrie, J. (2002). The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Statistics in Medicine*, *21*(19), 2899–2908. https://doi.org/10.1002/sim.1294

Ford, I., Norrie, J., & Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, *14*(8), 735–746. https://doi.org/10.1002/sim.4780140804

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67. https://doi.org/10.1214/aos/1176347963

Glynn, R. J., Gagne, J. J., & Schneeweiss, S. (2012). Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and Drug Safety*, *21*(S2), 138–147. https://doi.org/10.1002/pds.3231

Goemans, N., Wong, B., Van den Hauwe, M., Signorovitch, J., Sajeev, G., Cox, D., Landry, J., Jenkins, M., Dieye, I., Yao, Z., Hossain, I., Ward, S. J., & the Collaborative Trajectory Analysis Project (cTAP) (2020). Prognostic factors for changes in the timed 4-stair climb in patients with Duchenne muscular dystrophy, and implications for measuring drug efficacy: A multi-institutional collaboration. *PLOS One*, *15*(6). https://doi.org/10.1371/journal.pone.0232870

Hajage, D., De Rycke, Y., Chauvet, G., & Tubach, F. (2017). Estimation of conditional and marginal odds ratios using the prognostic score. *Statistics in Medicine*, *36*(4), 687–716. https://doi.org/10.1002/sim.7170

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481–488. https://doi.org/10.1093/biomet/asn004

Hernández, A. V., Steyerberg, E. W., & Habbema, J. D. F. (2004). Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology*, *57*(5), 454–460. https://doi.org/10.1016/j.jclinepi.2003.09.014

Herrera, A. F., Li, H., Castellino, S. M., Rutherford, S. C., Davison, K., Evans, A. G., Punnett, A., Constine, L. S., Hodgson, D. C., Parsons, S. K., Prica, A., Kostakoglu, L., Shipp, M. A., Laubach, C., Leblanc, M. L., Crump, M., Kahl, B. S., Leonard, J. P., Kelly, K. M., & Friedberg, J. W. (2020). SWOG S1826: A phase III, randomized study of Nivolumab plus AVD or Brentuximab Vedotin plus AVD in patients with newly diagnosed advanced stage classical Hodgkin lymphoma. *Blood*, *136*(Suppl. 1), 23–24. https://doi.org/10.1182/blood-2020-136422

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J. S., Selker, H. P., Varadhan, R., Vickers, A., Wong, J. B., & Steyerberg, E. W. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement: Explanation and elaboration. *Annals of Internal Medicine*, *172*(1), W1–W25. https://doi.org/10.7326/M18-3668

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M., & Horwitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*, *52*(1), 19–26. https://doi.org/10.1016/s0895-4356(98)00138-3

Kopp-Schneider, A., Calderazzo, S., & Wiesenfarth, M. (2020). Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. *Biometrical Journal*, *62*(2), 361–374. https://doi.org/10.1002/bimj.201800395

Kreuz, M., Otto, D. J., Fuessel, S., Blumert, C., Bertram, C., Bartsch, S., Loeffler, D., Puppel, S.-H., Rade, M., Buschmann, T., Christ, S., Erdmann, K., Friedrich, M., Froehner, M., Muders, M. H., Schreiber, S., Specht, M., Toma, M. I., Benigni, F., … Horn, F. (2020). Prostatrend—A multivariable prognostic RNA expression score for aggressive prostate cancer. *European Urology*, *78*(3), 452–459. https://doi.org/10.1016/j.eururo.2020.06.001

Lesaffre, E., & Senn, S. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine*, *22*(23), 3583–3596. https://doi.org/10.1002/sim.1583

Pfeiffer, R. M., & Riedl, R. (2015). On the use and misuse of scalar scores of confounders in design and analysis of observational studies. *Statistics in Medicine*, *34*(18), 2618–2635. https://doi.org/10.1002/sim.6467

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*, *21*(19), 2917–2930. https://doi.org/10.1002/sim.1296

Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, *31*, 172–181.

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale De Statistique*, *59*(2), 227–240. https://doi.org/10.1002/sim.1296

Schmid, M., Hothorn, T., Maloney, K. O., Weller, D. E., & Potapov, S. (2011). Geoadditive regression modeling of stream biological condition. *Environmental and Ecological Statistics*, *18*(4), 709–733. https://doi.org/10.1007/s10651-010-0158-4

Schuler, A., Walsh, D., Hall, D., Walsh, J., Fisher, C., & for the Critical Path for Alzheimer's Disease and the Alzheimer's Disease Neuroimaging Initiative and the Alzheimer's Disease Cooperative Study (2021). Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*. https://doi.org/10.1515/ijb-2021-0072

Steingrimsson, J. A., Hanley, D. F., & Rosenblum, M. (2017). Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemporary Clinical Trials*, *54*, 18–24. https://doi.org/10.1016/j.cct.2016.12.026

Wijdicks, E. F. M., Bamlet, W. R., Maramattom, B. V., Manno, E. M., & McClelland, R. L. (2005). Validation of a new coma scale: The FOUR score. *Annals of Neurology*, *58*(4), 585–593. https://doi.org/10.1002/ana.20611

Wyss, R., Glynn, R. J., & Gagne, J. J. (2016). A review of disease risk scores and their application in pharmacoepidemiology. *Current Epidemiology Reports*, *3*(4), 277–284. https://doi.org/10.1007/s40471-016-0088-2

Wyss, R., Lunt, M., Brookhart, M. A., Glynn, R. J., & Stürmer, T. (2014). Reducing bias amplification in the presence of unmeasured confounding through out-of-sample estimation strategies for the disease risk score. *Journal of Causal Inference*, *2*(2), 131–146. https://doi.org/10.1515/jci-2014-0009

Zhang, Z., & Ma, S. (2019). Machine learning methods for leveraging baseline covariate information to improve the efficiency of clinical trials. *Statistics in Medicine*, *38*(10), 1703–1714. https://doi.org/10.1002/sim.8054

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Siegfried, S., Senn, S., & Hothorn, T. (2022). On the relevance of prognostic information for clinical trials: A theoretical quantification. *Biometrical Journal*, 1–14. https://doi.org/10.1002/bimj.202100349

APPENDIX: ADJUSTING FOR COVARIATES: GAINS AND LOSSES

An easy way to see the effect of fitting covariates on the efficiency of an estimator is to consider adding a binary covariate (we shall take sex as an example) to the analysis of a design that is currently balanced by treatment with $2N$ patients per arm, there being two arms in total. If the covariate is not fitted, the variance of the treatment contrast will be

$$\left(\frac{1}{2N}+\frac{1}{2N}\right)\times\sigma_0^2=\frac{\sigma_0^2}{N},$$

where σ_0^2 is the within-treatment groups variance, which will be estimated using $2N-2$ degrees of freedom and where the subscript 0 is used to represent that no covariates have been fitted. Now suppose that the two sexes are equally well represented but having randomized and having decoded the data, we see that the disposition of subjects by group and sex is

	Control	Treatment	
Females	f	$2N-f$	$2N$
Males	$2N-f$	f	$2N$
	$2N$	$2N$	$4N=n$

where the entries in the cells represent frequencies of patients of the four types. The within-sex stratum estimates now have variances proportional to

$$\left(\frac{1}{f}+\frac{1}{2N-f}\right)\times\sigma_1^2,$$

where the subscript 1 is used to represent that one covariate has been fitted. Note, that one degree of freedom is lost if the fitting process uses sex as a main effect in an analysis of covariance. However, strict stratification estimates the variance within strata and loses one further degree of freedom. Here, we consider the former case, where the degrees of freedom available to estimate this variance are now $2N-3$. Clearly, the two within-stratum estimates are equally efficient and should be weighted equally, that is to say by one half. Thus, the combined estimate will have a variance equal to

$$\begin{aligned}\left(\frac{1}{2^2}\frac{1}{2^2}\right)\times\left(\frac{1}{f}+\frac{1}{2N-f}\right)\times\sigma_1^2&=\frac{1}{2}\times\left(\frac{2N}{f(2N-f)}\right)\times\sigma_1^2\\&=\frac{N}{f(2N-f)}\times\sigma_1^2.\end{aligned}$$

Note that the divisor of this expression can be expressed as $N^2-(f-N)^2$ and that $-(f-N)^2\leq 0$, so that the divisor reaches its maximum when $f=N$ that is to say the design is balanced, at which point, the variance will be σ_1^2/N.

We thus see that we can expect three consequences of fitting sex in the model. (1) If sex is predictive, we may expect $\sigma_1^2<\sigma_0^2$. We can refer to this as the *MSE effect*. (2) The variance multiplier will be

$$\frac{N}{f(2N-f)}\geq\frac{1}{N}$$

with equality only being achieved in the case of perfect balance. More generally, we may expect some imbalance and so some loss in efficiency. We can refer to this as the *imbalance effect*. (3) A completely predictable loss is that the degrees of freedom associated with the relevant t-distribution will be reduced by 1. This, unlike the other two effects, is not an effect on precision itself but an effect on our estimates of precision and may be referred to as the *second-order precision effect*. One way of judging it is to compare the variances of the two t-distributions involved, using the fact that in general this is $\nu/(\nu-2)$, where ν is the degrees of freedom. In the case with no predictors, we have $\nu = n-2$ and more generally, if we have k predictors, we have $\nu = n-2-k$ so that the general variance term is

$$\frac{n-2-k}{n-4-k},$$

with this reducing to $(n-2)/(n-4)$ if $k=0$, $(n-3)/(n-5)$ if $k=1$.

More generally, for the cases where covariates may be continuous and there may be more than one covariate but only two treatments, we may consider the influence of these three factors in terms of the general variance estimator $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sigma_k^2$. Here, $\boldsymbol{X}_{n\times(k+2)}$ is the design matrix for which we may assume, without loss of generality, that the first column is an intercept carrier, the second is a treatment indicator, and the k further columns, $k = 0, 1, 2, \dots$ are for the covariates.

This formulation includes not fitting covariates as a special case, for which $k = 0$. Note, however, that for the practical purpose of comparing using a single score based on covariates to using the original covariates themselves, then the lowest value that is of any interest is $k = 2$.

The diagonal elements of the $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ matrix give the variance multipliers and, given what we have said about the order of the columns, the second of these is the multiplier for the variance of the treatment effect. We refer to this as q_k, where the subscript k refers to the number of covariates being fitted and not to the position in the matrix . Thus, the variance of the treatment estimate is $q_k\sigma_k^2$. Given n patients, it can be shown that we must have $q_k \geq 4/n$. For our previous example, we had $n = 4N$, so we had $q_k \geq 1/N$.

As covariates are added to the model and therefore columns are added to the design matrix, the value of q_k cannot reduce but may increase. The example with sex as a binary covariate illustrates this. In a randomized design, the effect on q is not predictable as the design matrix will vary randomly but for normally distributed predictors the expected effect may be described. If the trial is balanced in the sense that there are the same number of patients on each of the two arms but otherwise randomized, the expected value is given by

$$\mathbb{E}(q_k) = \frac{4}{n}\times\frac{n-3}{n-3-k}.$$

Special cases are

$$\mathbb{E}(q_0) = \frac{4}{n}\times\frac{n-3}{n-3} = \frac{4}{n},$$
$$\mathbb{E}(q_1) = \frac{4}{n}\times\frac{n-3}{n-4}.$$

It thus follows that we have

$$\frac{\mathbb{E}(q_k)}{\mathbb{E}(q_0)} = \frac{n-3}{n-3-k},$$
$$\frac{\mathbb{E}(q_k)}{\mathbb{E}(q_1)} = \frac{n-4}{n-3-k},$$

the second of these being relevant to the task of comparing adjustment for a single score based on k covariates to independently fitting them all. Note that this formula does not depend on the covariates being generated by an independent process. (The covariates, could, for example, be correlated.) This is because, given k predictors and assuming that the set has no redundancy (the generating process is of rank k), they can be replaced by k orthogonal predictors, which together will have the same identical predictive value as the original k. Furthermore, if we have a predictive score, which is a linear combination of the predictors, then given $k-1$ predictors and the score, the value of the remaining predictor is completely determined and so redundant. Thus, the formula for $\mathbb{E}(q_k)/\mathbb{E}(q_1)$ is valid for this case also.

Thus, consider making a decision as to whether to fit such a score. A relevant comparison is that of the ratio of the two expected MSEs to the ratio of the expected imbalance factors. Thus, a sufficient condition for fitting such a score would be

$$\frac{\sigma_1^2}{\sigma_k^2} \leq \frac{n-4}{n-3-k}, \quad 2 \leq k \leq n-4,$$

where σ_1^2 is the MSE fitting the score as a single covariate. Note that the right-hand side of the expression is an expectation but a known quantity that must be greater than one (in expectation). The left-hand side is a random variable, which also ought to be greater than one, and some judgment must be made by the modeler as to what it will be. The lower bound of k is the lowest value of interest and the higher bound is the highest for which the expression on the right-hand side is defined.

One could also try to incorporate the second-order precision effect into the decision process. Note, however, that this is always in favor of using the score rather than the k individual predictors. Therefore, if the condition above is satisfied, it will definitely be an advantage to fit the score. This is why we refer to the condition as sufficient.

However, it should be noted, that the expression provides a means of guiding the choice between fitting k predictors and fitting a linear combination of them all. If $k \geq 3$ it is possible that fitting a reduced set would be better than either.