Biometrics, 2024, 80(1), ujad034
https://doi.org/10.1093/biomtc/ujad034
Reader Reaction

OXFORD

Adaptive selection of the optimal strategy to improve precision and power in randomized trials

Laura B. Balzer [1,*], Erica Cai[2], Lucas Godoy Garraza[3], Pracheta Amaranath[2]

[1]Division of Biostatistics, University of California Berkeley, Berkeley, CA 94720, United States, [2]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003, United States, [3]Department of Biostatistics, University of Massachusetts Amherst, Amherst, MA 01003, United States

[*]Corresponding author: Laura B. Balzer, Division of Biostatistics, University of California Berkeley, Berkeley, CA, 94720, United States (laura.balzer@berkeley.edu).

ABSTRACT

Benkeser et al. demonstrate how adjustment for baseline covariates in randomized trials can meaningfully improve precision for a variety of outcome types. Their findings build on a long history, starting in 1932 with R.A. Fisher and including more recent endorsements by the U.S. Food and Drug Administration and the European Medicines Agency. Here, we address an important practical consideration: how to select the adjustment approach—which variables and in which form—to maximize precision, while maintaining Type-I error control. Balzer et al. previously proposed Adaptive Pre-specification within TMLE to flexibly and automatically select, from a prespecified set, the approach that maximizes empirical efficiency in small trials ($N < 40$). To avoid overfitting with few randomized units, selection was previously limited to working generalized linear models, adjusting for a single covariate. Now, we tailor Adaptive Pre-specification to trials with many randomized units. Using V-fold cross-validation and the estimated influence curve-squared as the loss function, we select from an expanded set of candidates, including modern machine learning methods adjusting for multiple covariates. As assessed in simulations exploring a variety of data-generating processes, our approach maintains Type-I error control (under the null) and offers substantial gains in precision—equivalent to 20%-43% reductions in sample size for the same statistical power. When applied to real data from ACTG Study 175, we also see meaningful efficiency improvements overall and within subgroups.

KEYWORDS: covariate adjustment; efficiency; machine learning; pre-specification; randomized trials; TMLE.

1 INTRODUCTION

There is a long history of debate on whether and how to optimally adjust for baseline covariates to improve precision in randomized trials (eg, Fisher (1932); Tsiatis et al. (2008); Zhang et al. (2008); EMA (2015); FDA (2021)). Recently, for binary, ordinal, and time-to-event outcomes, Benkeser et al. (2021) defined several potential causal effects of interest and, for each, demonstrated the promise of covariate adjustment to improve our ability to make timely and precise inferences, without fear of bias due to regression model misspecification. In their simulation study, covariate adjustment led to substantial gains in efficiency, translating to 4%-18% reductions in sample size for the same statistical power when there was an effect, while maintaining good Type-I error control when there was no effect.

However, Benkeser et al. (2021) only briefly discuss how to optimally select the adjustment covariates and the "working" regression model to maximize precision for the effect of interest. In their Rejoinder, they state, "the variables should either be selected before the trial starts (selecting those that are most prognostic for the outcome based on prior data), or selected using the trial data based on a completely prespecified algorithm that aims to select the most prognostic variables" (Benkeser et al., 2021). Their recommendation to use prior data assumes the existence of such data and the consistency of relationships over time and space. Their recommendation for data-adaptive selection does not provide a specific algorithm.

The challenge of "how" is further highlighted in the corresponding commentaries. Specifically, Zhang and Zhang (2021) emphasize that using a misspecified regression model for covariate adjustment can improve efficiency "as long as the covariates are predictive of the outcomes". This leaves the reader wondering about the potential detriments to efficiency and Type-I error control with forced adjustment for covariates that are, in fact, not prognostic of the outcome. Zhang and Zhang (2021) call for further investigation into practical challenges, such as few independent units, stratified randomization, and covariate selection.

Building on our previous work in Adaptive Pre-specification (APS), we offer concrete solutions to these practical challenges (Balzer et al., 2016). Our approach data-adaptively selects, from a prespecified set, the covariates and the form of the working model to minimize the cross-validated variance estimate and, thereby, maximize the empirical efficiency. Our approach is applicable to asymptotically linear estimators with known influence curves and, thus, covers a large class of algorithms including those most commonly used for causal inference. Additionally, our work is applicable under a variety of trial designs:

Received: October 31, 2022; Revised: September 6, 2023; Accepted: December 15, 2023

simple randomization, randomization within strata of baseline covariates, and randomization within matched pairs of units. Throughout, we focus on a targeted minimum-loss based estimation (TMLE), which is summarized below and detailed in Web Appendix A (van der Laan and Rose, 2011). However, we emphasize APS can be applied to select the optimal, asymptotically linear estimator for a wide variety of causal effects.

2 METHODS

In randomized trials, TMLE is a powerful approach to leverage baseline covariates for efficiency gains, while remaining robust to regression model misspecification (eg, Moore and van der Laan (2009); Rosenblum and van der Laan (2010); Balzer et al. (2023); Benitez et al. (2023)). For outcomes measured completely in a 2-armed trial, the steps of TMLE are (1) obtain an initial estimator of the "outcome regression", defined as the conditional expectation of the outcome Y given the treatment indicator A and covariates W; (2) obtain predicted outcomes under the treatment $\hat{\mathbb{E}}(Y|A=1,W)$ and under the control $\hat{\mathbb{E}}(Y|A=0,W)$; (3) target these outcome predictions using information in the "propensity score", defined as the conditional probability of receiving the treatment given the covariates $\mathbb{P}(A=1|W)$; (4) average the targeted predictions under the treatment $\hat{\mathbb{E}}^*(Y|A=1,W)$, and under the control $\hat{\mathbb{E}}^*(Y|A=0,W)$; and (5) contrast on the scale of interest.

Under standard regularity conditions, TMLE is asymptotically linear; therefore, the standardized estimator is asymptotically normal with mean zero and variance given by the variance of its influence curve (Web Appendix A). If the initial estimator for the outcome regression $\mathbb{E}(Y|A,W)$ uses a "working" generalized linear model (GLM) with an intercept and a main term for the treatment and if the propensity score is not estimated, then targeting (step 3) can be skipped (Rosenblum and van der Laan, 2010). Further precision can be attained by using a prespecified, data-adaptive algorithm for initial estimation of the outcome regression and for "collaborative" estimation of the propensity score, as described next.

2.1 Adaptive Pre-specification (APS)

Rubin and van der Laan (2008) proposed the principle of empirical efficiency maximization to optimize precision and, thus, power in randomized trials. To pick the "best" estimator, they propose using as loss function the squared-efficient influence curve for the estimand of interest; this corresponds to selecting the candidate TMLE with the smallest cross-validated variance estimate (van der Laan and Rose (2011); pp. 572-577). Building on this principle and motivated by the SEARCH community randomized trial ($N=32$), Balzer et al. (2016) proposed and implemented APS to select the optimal adjustment approach in trials with few randomized units. We now generalize APS for use in trials with many randomized units.

Figure 1 provides a schematic of APS to select and implement the optimal TMLE (details in Web Appendix A). First, we prespecify candidate estimators of the outcome regression and of the known propensity score. Together, they form the set of candidate TMLEs, which are consistent and locally efficient estimators. As originally formulated for small trials, we limited candidate estimators of the outcome regression to working GLMs with an intercept, a main term for the treatment A, and at most 1 adjustment variable. We also limited candidate estimators of the propensity score to working logistic regressions with an intercept and at most 1 adjustment variable. For the large-trial implementation, we now propose adjusting for multiple covariates in more flexible algorithms, such as penalized regression and multivariate adaptive regression splines (MARS). The unadjusted estimator must also be included as a candidate. APS also requires us to prespecify a cross-validation (CV) scheme and a loss function to objectively measure performance. For small trials, we used leave-one-out CV; for larger trials, we use V-fold CV. As loss function, we use the estimated influence curve-squared for the TMLE of the effect of interest (Web Appendix A).

With these ingredients, we have a fully prespecified and automated procedure to data-adaptively select the TMLE that maximizes empirical efficiency. Selection occurs in 2 steps. First, we select the candidate estimator of the outcome regression—both the adjustment variable(s) and the functional form—that minimizes the cross-validated variance estimate. Next, we select the candidate estimator of the propensity score—both the adjustment variable(s) and the functional form—that further minimizes the cross-validated variance estimate when used to target initial predictions from the previously selected outcome regression estimator. The two selected estimators form the optimal TMLE, which is then fit using all the data.

APS can be considered an extension of Collaborative-TMLE using a CV-selector (a.k.a., discrete Super Learner) to maximize precision in randomized trials. In the small trial setting, substantial precision gains from APS have repeatedly been demonstrated in simulations and with real data (Balzer et al., 2016; Benitez et al., 2023). For example, in the SEARCH study ($N=32$), we found that the variance of the unadjusted effect estimator was 4.6 times that of TMLE with the small-trial implementation of APS (Balzer et al., 2023). We now examine the performance of our proposed modification to APS for large-trial settings using simulations as well as a real-data application.

3 SIMULATION STUDIES

To address the persistent concerns about adjustment with nonlinear models, highlighted by LaVange (2021) among others, we explore data-generating processes with higher order interactions and nonlinear link functions. We also evaluate the performance with simple randomization and stratified randomization. Our focus is on estimating the sample effect; however, as previously discussed, our approach is applicable to other asymptotically linear estimators with known influence curves, such as TMLE for the population average treatment effect (PATE) or the conditional average treatment effect (CATE; Web Appendix A).

We consider 5000 simulated trials, each with $N=500$ participants. For each participant, we generate 5 measured covariates $\{W_1,\ldots,W_5\}$ from a standard normal distribution, 2 unmeasured covariates $\{U_1,U_2\}$ from a standard uniform distribution, and the binary counterfactual outcomes $Y(a)$ in 3 settings of varying complexity:

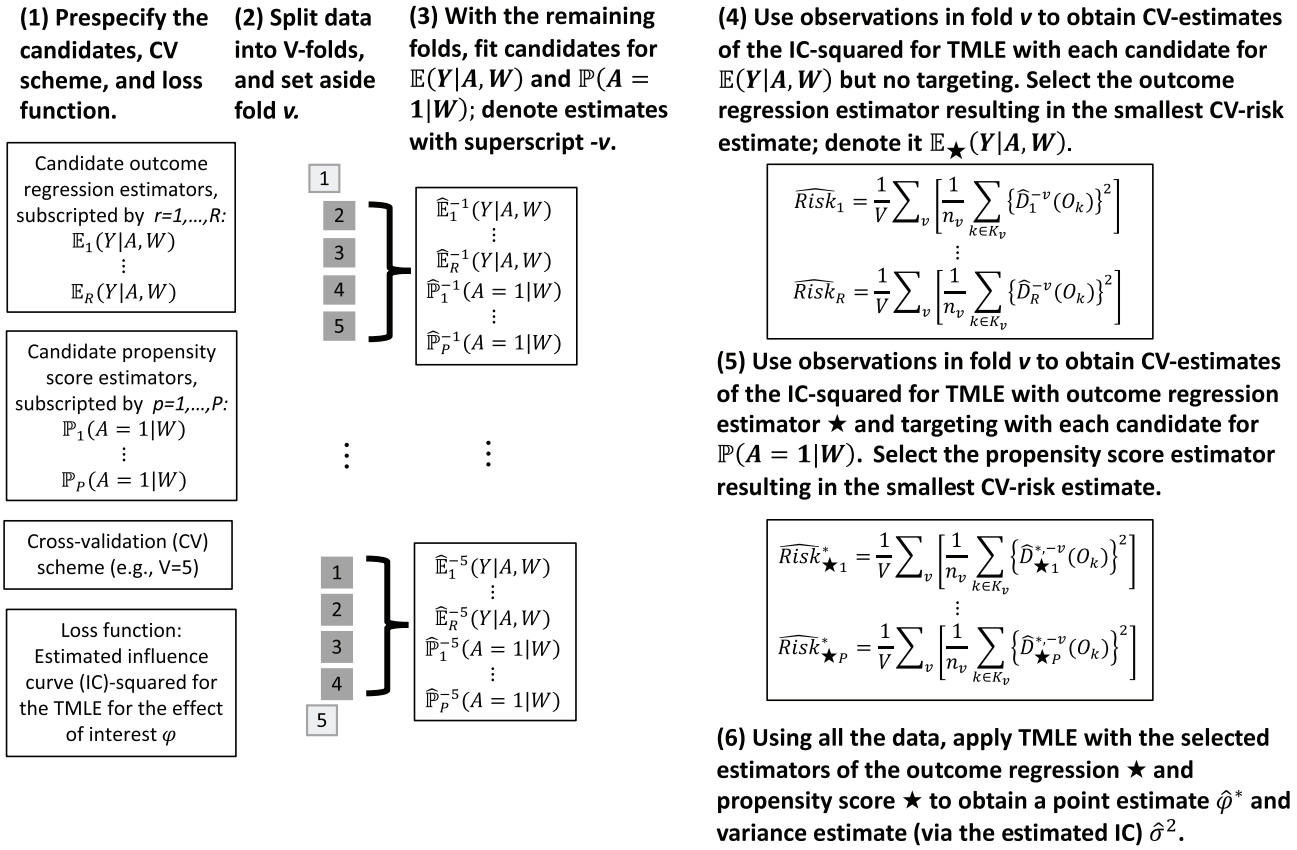

FIGURE 1 Schematic of Adaptive Pre-specification (APS) within TMLE to flexibly and automatically select, from a prespecified set, the adjustment approach that maximizes empirical efficiency for the effect of interest. For illustration, we show R candidate outcome regression estimators $\mathbb{E}(Y|A,W)$, P candidate propensity score estimators $\mathbb{P}(A=1|W)$, and $V=5$-fold cross-validation (CV). For simplicity, we show the process for first and last folds, and use ellipses to indicate an analogous process for the other folds. Let K_v denote the set of indices for the observations in fold v of size $|K_v|=n_v$. For observation k in validation set v, the CV-influence curve estimate for the TMLE using candidate outcome regression r but no targeting is denoted $\hat{D}_r^{-v}(O_k)$ in step 4, while the corresponding CV-estimate of the influence curve for the TMLE using the selected outcome regression $\star$ and targeting with candidate propensity score estimator p is denoted $\hat{D}^{*,-v}_{\star p}(O_k)$ in step 5 (Web Appendix A).

(i) "linear": $Y(a)=\mathbb{1}\{U_1 < logit^{-1}(a+W_1-W_2+W_3-W_4+W_5-2aW_1+U_2)\}$,
(ii) "interactive": $Y(a)=\mathbb{1}\{U_1 < logit^{-1}(a+W_1+W_2+W_3+W_4+W_5+aW_1+aW_2W_4+aW_3+aW_5U_2+U_2)\}$,
(iii) "polynomial": $Y(a)=\mathbb{1}\{U_1 < logit^{-1}(a+W_1+W_2+W_3+W_4+W_5-W_1W_3+2W_1W_3W_4-W_4(1-W_1)+U_2)\}$.

We additionally consider a "treatment only" scenario where none of the measured covariates influences the outcome: $Y(a)=\mathbb{1}\{U_1 < logit^{-1}(0.1a+2U_2)\}$. Using these counterfactual outcomes, we calculate the true value of the sample risk ratio $(RR)=1/N\sum_i Y_i(1) \div 1/N\sum_i Y_i(0)$. As detailed in Web Appendix B, we also consider a continuous outcome, generated under 4 analogous settings. For the continuous endpoint, we focus on the sample average treatment effect $(SATE)=1/N\sum_i Y_i(1) - 1/N\sum_i Y_i(0)$. In each setting, we generate the observed treatment A using simple randomization and randomization within strata defined by $\mathbb{1}(W_1>0)$. Finally, we set the observed outcome Y equal to the counterfactual outcome $Y(a)$ when the observed treatment $A=a$.

We compare the unadjusted estimator, fixed adjustment for W_1 in the outcome regression, TMLE with APS tailored for small trials, and TMLE with our novel modification of APS for larger trials. In the small-trial APS, we limit the candidate estimator to working GLMs with 1 adjustment covariate selected from $\{W_1, W_2, W_3, W_4, W_5, \emptyset\}$. In the large-trial APS, we select from working GLMs adjusting for 1 covariate, main terms GLM adjusting for all covariates, stepwise regression, stepwise regression with all possible pairwise interactions, LASSO, MARS, and the unadjusted estimator. Both versions of APS use 5-fold CV. Performance criteria include 95% confidence interval coverage, attained power, Type-I error (under the null), bias, variance, and mean squared error (MSE). Following Benkeser et al. (2021), we provide the relative efficiency, calculated as the MSE of a covariate-adjusted estimator divided by the MSE of the unadjusted effect estimator, and provide an estimate the

TABLE 1 Estimator performance with the binary outcome where there is an effect and with a sample size of $N = 500$.

DGP	Design	Estimator	Cover.	Power	MSE	Bias	Var.	Rel.Eff.
Linear	Simple	Unadjusted	0.976	0.929	0.005	0.002	0.007	1.000
		Static	0.977	0.929	0.005	0.002	0.007	0.998
		Small APS	0.982	0.950	0.004	− 0.006	0.006	0.798
		Large APS	0.971	0.984	0.004	0.002	0.006	0.699
	Stratified	Unadjusted	0.980	0.935	0.005	0.003	0.007	1.000
		Static	0.979	0.935	0.005	0.003	0.007	0.997
		Small APS	0.984	0.952	0.004	− 0.005	0.006	0.796
		Large APS	0.974	0.984	0.004	0.003	0.006	0.707
Interactive	Simple	Unadjusted	0.965	0.211	0.003	0.003	0.003	1.000
		Static	0.964	0.228	0.002	0.003	0.003	0.867
		Small APS	0.975	0.212	0.002	0.001	0.002	0.730
		Large APS	0.965	0.324	0.002	0.003	0.002	0.567
	Stratified	Unadjusted	0.970	0.184	0.003	0.001	0.003	1.000
		Static	0.965	0.216	0.002	0.001	0.003	0.945
		Small APS	0.980	0.190	0.002	− 0.001	0.002	0.774
		Large APS	0.966	0.297	0.002	0.001	0.002	0.601
Polynomial	Simple	Unadjusted	0.963	0.865	0.004	0.004	0.005	1.000
		Static	0.967	0.916	0.004	0.002	0.004	0.811
		Small APS	0.970	0.914	0.003	− 0.004	0.004	0.719
		Large APS	0.969	0.969	0.003	0.001	0.003	0.584
	Stratified	Unadjusted	0.977	0.868	0.004	0.002	0.004	1.000
		Static	0.966	0.912	0.003	0.002	0.004	0.917
		Small APS	0.974	0.908	0.003	− 0.004	0.004	0.816
		Large APS	0.972	0.972	0.003	0.001	0.003	0.673

Note: "DGP" denotes the data-generating process; "Cover." denotes the 95% confidence interval coverage; "Power" denotes the proportion of times the true null hypothesis was rejected; "MSE" denotes mean squared error; "Var." denotes the variance of the point estimates, and "Rel.Eff." denotes relative efficiency, approximated by the ratio of the MSE of a given estimator to that of the unadjusted estimator. The average value of the sample risk ratio is 1.25 in the linear setting, 1.06 in the interactive setting, and 1.19 in the polynomial setting. "Static" refers to forced adjustment for W_1 in the outcome regression, "Small APS" to TMLE with the small-trial implementation of Adaptive Pre-specification (APS), and "Large APS" to TMLE with the large-trial implementation of APS.

potential savings in sample size, calculated as 1 minus the relative efficiency.

For the binary outcome, TMLE using the large-trial APS substantially improved power for the risk ratio in all scenarios (Table 1). Absolute gains in power as compared to the unadjusted approach ranged from 5% to 11%. The relative efficiency ranged from 0.57 to 0.71. This roughly translates to 29%-43% savings in sample size from using TMLE, instead of the unadjusted effect estimator (Figure 2). The gains from TMLE with the small-trial APS were less extreme, but still notable (relative efficiency: 0.72-0.82). Importantly, all estimators maintained nominal-to-conservative confidence interval coverage, as expected when estimating sample effects (Web Appendix A).

Estimator performance with a continuous outcome and targeting SATE was similar (Table 2). In all scenarios, TMLE with the large-trial APS achieved the highest power with absolute gains of 18%-23% compared to the unadjusted estimator. Its relative efficiency was 0.58-0.80, roughly translating to 20%–42% savings in sample size (Figure 2). As before, the gains from TMLE with the small-trial APS were less extreme, but still notable (relative efficiency: 0.60–0.98). Again, the 95% confidence interval coverage of the adaptive estimators was comparable to that of the unadjusted estimator.

Additional results are available in Web Appendix B. The selection of estimators for the outcome regression and propensity score varied by setting, highlighting our approach's ability to respond to the data-generating process (Web Tables 1-2). Importantly, for both outcome types, the precision gains from TMLE were achieved without sacrificing Type-I error, even in the "treatment only" setting where there were no prognostic covariates (Web Tables 3-4).

4 REAL DATA APPLICATION: ACTG STUDY 175

ACTG Study 175 evaluated the effect of monotherapy versus combination therapy on health outcomes among persons with HIV (Hammer et al. 1996). For demonstration, we examine the effect of the antiretroviral therapy (ART) regimen only containing zidovudine ($A = 0$) versus on alternative regimen ($A = 1$) on the difference in the average CD4 count at 20 weeks (continuous outcome) and the relative risk of the 20-week CD4 count>350c/mm^3 (binary outcome). We use the unadjusted estimator, fixed adjustment for age and gender, TMLE with the small-trial APS, and TMLE with the large-trial APS. Within APS, we considered 16 candidate adjustment variables, including demographics and ART history (Web Table 5).

The results are summarized in Table 3, with further details in Web Appendix C. As expected, the point estimates were similar across approaches, but the application of APS offered notable precision gains. The estimated variance of TMLE with the large-trial APS divided by that of the unadjusted approach was 0.54 and 0.67 for the continuous and binary outcome, respectively. Assuming negligible bias, this would roughly translate into needing 46% and 33% fewer participants with our approach. Importantly, Type-I error control, evaluated through treatment-blind simulations, was maintained at the nominal rate of 5%. As

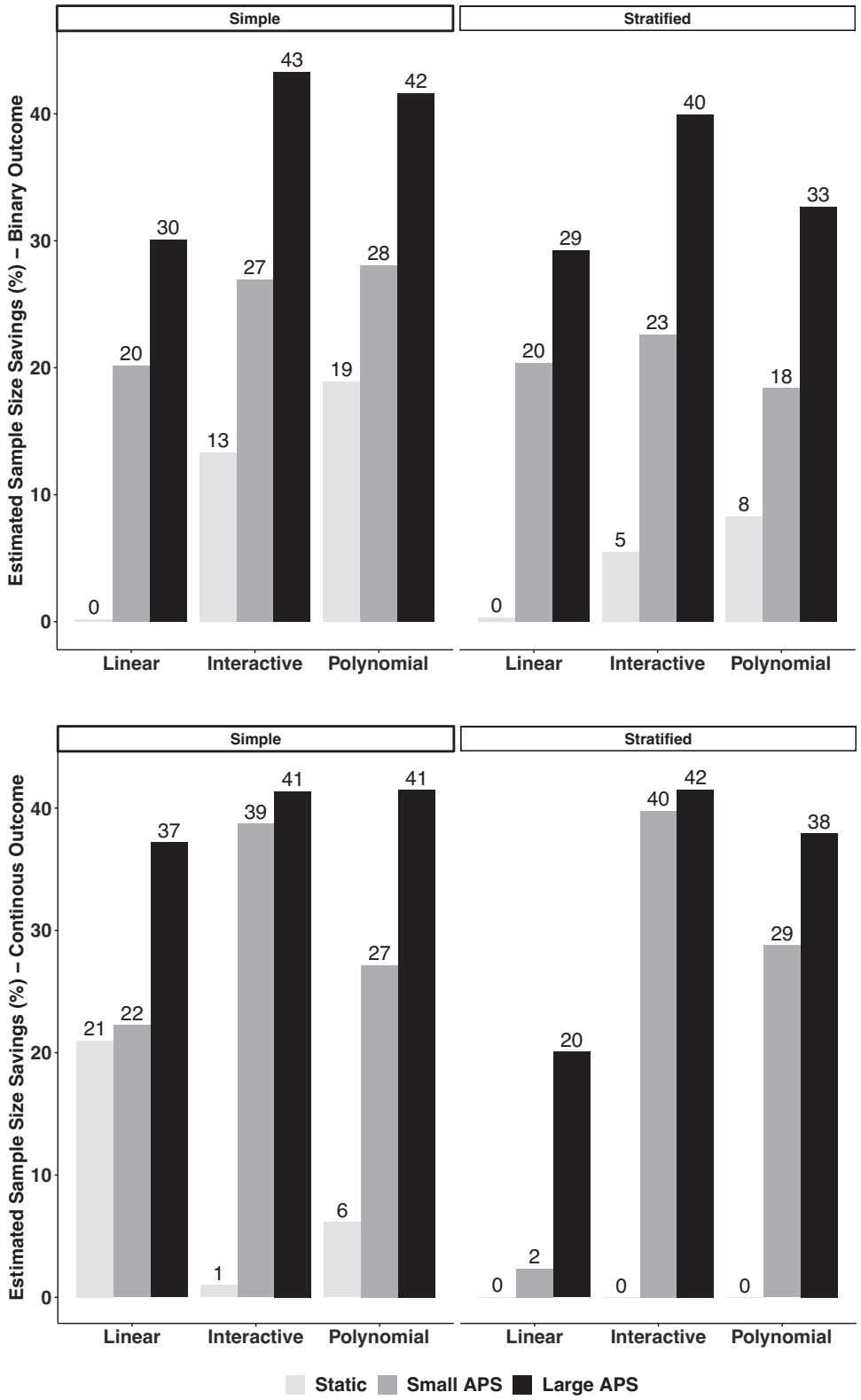

FIGURE 2 Across 5000 simulated trials with a binary outcome (top) and with a continuous outcome (bottom), the estimated savings in sample size (in %), as compared to the unadjusted estimator, when using forced adjustment for W_1 in the outcome regression ("Static"), TMLE with the small-trial implementation of Adaptive Pre-specification ("Small APS"), and TMLE with the large-trial implementation ("Large APS") across the 3 data-generating processes with prognostic covariates and with simple versus stratified randomization.

TABLE 2 Estimator performance with the continuous outcome where there is an effect and with a sample size of $N = 500$.

DGP	Design	Estimator	Cover.	Power	MSE	Bias	Var.	Rel.Eff.
Linear	Simple	Unadjusted	0.953	0.591	0.008	−0.001	0.008	1.000
		Static	0.948	0.692	0.006	−0.002	0.006	0.790
		Small APS	0.950	0.689	0.006	−0.002	0.006	0.778
		Large APS	0.950	0.794	0.005	−0.001	0.005	0.628
	Stratified	Unadjusted	0.971	0.617	0.006	0.001	0.006	1.000
		Static	0.945	0.705	0.006	0.001	0.006	1.000
		Small APS	0.947	0.706	0.006	0.000	0.006	0.977
		Large APS	0.944	0.799	0.005	0.001	0.005	0.799
Interactive	Simple	Unadjusted	0.948	0.389	0.011	−0.001	0.011	1.000
		Static	0.946	0.399	0.011	−0.001	0.011	0.990
		Small APS	0.948	0.578	0.007	−0.001	0.007	0.613
		Large APS	0.946	0.602	0.007	−0.001	0.007	0.586
	Stratified	Unadjusted	0.942	0.402	0.012	0.003	0.012	1.000
		Static	0.939	0.409	0.012	0.003	0.012	1.000
		Small APS	0.946	0.586	0.007	0.001	0.007	0.602
		Large APS	0.943	0.617	0.007	0.001	0.007	0.585
Polynomial	Simple	Unadjusted	0.948	0.489	0.008	−0.001	0.008	1.000
		Static	0.943	0.518	0.007	−0.001	0.007	0.938
		Small APS	0.953	0.613	0.006	−0.002	0.006	0.729
		Large APS	0.948	0.724	0.004	−0.001	0.004	0.585
	Stratified	Unadjusted	0.948	0.508	0.007	0.002	0.007	1.000
		Static	0.939	0.536	0.007	0.002	0.007	1.000
		Small APS	0.957	0.631	0.005	0.001	0.005	0.712
		Large APS	0.942	0.735	0.005	0.001	0.005	0.621

Note. "DGP" denotes the data-generating process; "Cover." denotes the 95% confidence interval coverage; "Power" denotes the proportion of times the true null hypothesis was rejected; "MSE" denotes mean squared error; "Var." denotes the variance of the point estimates, and "Rel.Eff." denotes relative efficiency, approximated by the ratio of the MSE of a given estimator to that of the unadjusted estimator. The average value of the sample average treatment effect (SATE) is 0.195 in the linear setting, 0.180 in the interactive setting, and 0.170 in the polynomial setting. "Static" refers to forced adjustment for W_1 in the outcome regression, "Small APS" to TMLE with the small-trial implementation of Adaptive Pre-specification (APS), and "Large APS" to TMLE with the large-trial implementation of APS.

TABLE 3 Comparative results using real data from ACTG Study 175 to estimate the effect on difference in the average CD4 count at 20 weeks (continuous outcome) and on the relative risk of 20-week CD4 count $>350c/mm^3$ (binary outcome).

Outcome	Estimator	Effect (95%CI)	Rel.Var.	Out.Reg.	PScore	Type-I
Continuous	Unadjusted	46.4 (33.0, 59.7)	1.000	Unadj.	Unadj.	4.8%
	Static	46.8 (33.5, 60.0)	0.991	Fixed	Fixed	4.8%
	Small APS	48.5 (38.0, 59.0)	0.617	GLM	GLM	5.2%
	Large APS	47.8 (38.0, 57.6)	0.542	MARS	GLM	5.3%
Binary	Unadjusted	1.23 (1.10, 1.37)	1.000	Unadj.	Unadj.	4.7%
	Static	1.23 (1.11, 1.37)	1.001	Fixed	Fixed	4.5%
	Small APS	1.26 (1.15, 1.38)	0.702	GLM	GLM	5.2%
	Large APS	1.26 (1.15, 1.37)	0.672	LASSO	GLM	5.2%

Note. "Static" refers to forced adjustment for age in the outcome regression and gender in the propensity score, "Small APS" to TMLE with the small-trial implementation of Adaptive Pre-specification (APS), selecting from working GLMs adjusting for at most 1 covariate, and "Large APS" to TMLE with the large-trial implementation of APS, selecting from the small-trial algorithms, main terms, stepwise regression, LASSO, MARS, and MARS after correlation-based screening. "Rel.Var." is the estimated variance of a given approach divided by the estimated variance of the unadjusted approach. "Out.Reg." is the selected approach for estimation of the outcome regression, and "PScore" is the selected approach for estimation of the known propensity score. "GLM" refers to a working GLM adjusted for at most 1 covariate. "Type-I" is the estimated Type-I error rate, evaluated with treatment-blind simulations, which permute the treatment indicator A, implement each estimator, and repeat 5000 times.
Additional details and results are given in Web Appendix C.

expected, the optimal TMLE varied by the target of inference and sample size (Web Tables 6–7). For smaller subgroups of older and younger women, there were notable gains in efficiency from estimation of the propensity score, but no difference between the TMLEs using the large versus small-trial APS. Here, both APS implementations selected a working GLM adjusting for 1 covariate when estimating the outcome regression and when estimating the propensity score. In contrast, overall and for larger subgroups of older and younger men, TMLE with the large-trial APS offered notable precision gains over the small-trial implementation, but there were minimal precision improvements from propensity score estimation.

5 DISCUSSION

The U.S. Food and Drug Administration and the European Medicines Agency endorse adjustment for baseline covariates to improve precision and, thereby, power in randomized trials

(EMA, 2015; FDA, 2021). Nonetheless, explicit guidance on how to optimally select and incorporate adjustment variables has been lacking. Indeed, the challenges in practical implementation were discussed in Benkeser et al. (2021) and the accompanying commentaries. For trials with limited numbers of randomized units ($N < 40$), Balzer et al. (2016) addressed this gap with APS, which selects the adjustment strategy maximizing the empirical efficiency. Here, for trials with many randomized units, we extended APS to include machine learning algorithms (eg, LASSO and MARS) adjusting for multiple covariates. Our simulations demonstrated improved precision and power, translating to 18%-43% potential savings in sample size, while controlling Type-I error. These gains were seen across a variety of data-generating processes, for both binary and continuous outcomes, and for both absolute and relative effects. Our real data application also demonstrated precision gains and highlighted how selection of the optimal TMLE was responsive to the subgroup.

Our approach offers several advantages over other model-robust, covariate-adjusted estimators. First, it is estimand-aligned; we can estimate user-specified effects on any scale (eg, difference, ratio), for any inferential target (eg, sample, conditional, or population effect), for several study designs (eg, simple, stratified, matched), and for a variety of outcome types (eg, binary, continuous). Second, our approach is fully prespecified, while remaining data-adaptive. Practically, we can prespecify several candidate estimators of the outcome regression, and let the algorithm pick the best approach, where "best" means maximizing empirical efficiency. These candidates can include user-specified GLMs, including known or suspected interactions, as well as modern advances in machine learning. Third, our approach incorporates collaborative estimation of the known propensity score for additional gains in precision. Thereby, we only estimate $\mathbb{P}(A = 1|W)$ if it improves the empirical efficiency; otherwise, we treat the propensity score as known and only adjust in the outcome regression. Collaborative estimation of the propensity score does come at the cost of a more complicated algorithm. However, it can meaningfully improve precision, especially in small trials (Balzer et al., 2016) or smaller subgroups (Web Tables 6-7), and computing code is readily available. Finally, if we are in the unfortunate scenario where adjustment does not improve efficiency, the algorithm will default to the unadjusted effect estimator. Thus, we are protected from forced adjustment at the detriment of precision or Type-I error control.

We have the following recommendations when implementing APS. First, increase the number of cross-validation (CV) folds as the number of randomized units decreases. Second, as candidates, consider a diverse set of asymptotically linear estimators. To prevent forced adjustment when harmful to precision, always include the unadjusted estimator as a candidate. As shown here, including candidates that flexibly adjust for multiple covariates, while satisfying the usual regularity conditions, can lead to substantial savings in sample size (Figure 2). If considering more aggressive algorithms (eg, random forest) that do not readily satisfy the conditions for asymptotic linearity, additional sample splitting is recommended. APS naturally generates a cross-validated variance estimate, which can be used if there are concerns about overfitting. To guide development of the Statistical Analysis Plan, we strongly recommend conducting a simulation study, reflecting the real data application, to facilitate prespecification of the candidate estimators, the CV scheme, and the variance estimator based on objective criteria (eg, relative efficiency and Type-I error control).

There are several limitations to our presentation. First, we focused on using APS to chose between candidate TMLEs; however, the procedure is applicable to other asymptotically linear estimators with known influence curves. This includes more traditional estimators, such as the Cox model for time-to-event outcomes. We plan to apply APS to select the optimal approach from a variety of doubly robust candidates, such as TMLE, augmented inverse probability weighting, and double/debiased machine learning. Second, we focused on trials where outcomes were measured completely; our work is immediately applicable to settings where censoring or missingness is completely at random. If, instead, censoring or missingness is random conditional on a X, a subset of the full covariate set W, APS should also be applicable with the following modification: all candidates must adjust for X and may consider additional adjustment for the remaining covariates. Further investigation is warranted. (We refer the reader to Balzer et al. (2023) for the application of APS in cluster randomized trials with missing or censored outcomes.) Third, our approach is applicable to trials with simple randomization, stratified randomization, and randomization within matched pairs; further investigation is needed for settings with sequential randomization. Finally, as currently implemented, APS uses a CV-selector (a.k.a., discrete Super Learner) to choose the single best estimator of the outcome regression combined with the single best estimator of the propensity score. We are working to extend APS to select the optimal convex combination of candidate estimators. Nonetheless, we believe TMLE with APS, as currently implemented, is a powerful and under-utilized tool for optimal covariate adjustment in randomized trials.

ACKNOWLEDGMENTS

We thank Drs. Alan Hubbard, Maya Petersen, and Mark van der Laan for their feedback.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices, referenced in Sections 2-4, are available with this paper at the *Biometrics* website on Oxford Academic. Appendix A provides details on TMLE with and without Adaptive Pre-specification (APS). Appendix B and C provide additional details and results for the simulation studies and real data application, respectively.

FUNDING

This work was supported, in part, by the National Institutes of Health (NIH; U01AI150510, R01AI074345), and DARPA (HR001120C0031). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the

authors and do not necessarily reflect the views of the NIH or the United States Air Force.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data that support the findings of this paper are openly available in GitHub at https://github.com/LauraBalzer/AdaptivePrespec. Computing code, including a vignette with worked examples, is also available via this GitHub.

REFERENCES

Balzer, L., van der Laan, M., Ayieko, J., Kamya, M., Chamie, G., Schwab, J. et al. (2023). Two-stage TMLE to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics*, 24, 502–517.

Balzer, L., van der Laan, M., Petersen, M. and SEARCH (2016). Adaptive pre-specification in randomized trials with and without pair-matching. *Stat Med*, 35, 4528–4545.

Benitez, A., Petersen, M., van der Laan, M., Santos, N., Butrick, E., Walker, D. M. et al. (2023). Comparative methods for the analysis of cluster randomized trials. *Stat Med*, 42, 3443–3466.

Benkeser, D., Díaz, 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 (with rejoinder). *Biometrics*, 77, 1467–1481.

EMA (2015). Guideline on adjustment for baseline covariates in clinical trials. https://www.ema.europa.eu/documents/scientific-guideline/guideline-adjustment-baseline-covariates-clinical-trials_en.pdf.

FDA (2021). COVID-19: Developing drugs and biological products for treatment or prevention. Guidance for industry. https://www.fda.gov/media/137926/download.

Fisher, R. (1932). *Statistical Methods for Research Workers*, Oliver and Boyd Ltd.: Edinburgh.

Hammer, S., Katzenstein, D., Hughes, M., Gundacker, H., Schooley, R., Haubrich, R. et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per mm^3. *NEJM*, 335, 1081–1090.

LaVange, L. (2021). Discussion of "improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes". *Biometrics*, 77, 1489–1491.

Moore, K. and van der Laan, M. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Stat Med*, 28, 39–64.

Rosenblum, M. and van der Laan, M. (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *Int J Biostat*, 6, 13.

Rubin, D. and van der Laan, M. (2008). Empirical efficiency maximization. *Int J Biostat*, 4, 5.

Tsiatis, A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials. *Stat Med*, 27, 4658–4677.

van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer: New York Dordrecht Heidelberg London.

Zhang, M., Tsiatis, A. and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates (with rejoinder). *Biometrics*, 64, 707–715.

Zhang, M. and Zhang, B. (2021). Discussion of "improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes". *Biometrics*, 77, 1485–1488.

Received: October 31, 2022; Revised: September 6, 2023; Accepted: December 15, 2023