Title
A generalization of moderated statistics to data adaptive semiparametric estimation in high-dimensional biology

Journal
Statistical Methods in Medical Research, 32(3)

ISSN
0962-2802

Authors
Hejazi, Nima S
Boileau, Philippe
van der Laan, Mark J
et al.

Publication Date
2023-03-01

DOI
10.1177/09622802221146313

Peer reviewed

HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2024 May 08.

Published in final edited form as:

Stat Methods Med Res. 2023 March ; 32(3): 539–554. doi:10.1177/09622802221146313.

A generalization of moderated statistics to data adaptive semiparametric estimation in high-dimensional biology

Nima S Hejazi[1], **Philippe Boileau**[2,3], **Mark J van der Laan**[2,3,4], **Alan E Hubbard**[2,3]

[1]Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

[2]Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA

[3]Center for Computational Biology, University of California, Berkeley, CA, USA

[4]Department of Statistics, University of California, Berkeley, CA, USA

Abstract

The widespread availability of high-dimensional biological data has made the simultaneous screening of many biological characteristics a central problem in computational and high-dimensional biology. As the dimensionality of datasets continues to grow, so too does the complexity of identifying biomarkers linked to exposure patterns. The statistical analysis of such data often relies upon parametric modeling assumptions motivated by convenience, inviting opportunities for model misspecification. While estimation frameworks incorporating flexible, data adaptive regression strategies can mitigate this, their standard variance estimators are often unstable in high-dimensional settings, resulting in inflated Type-I error even after standard multiple testing corrections. We adapt a shrinkage approach compatible with parametric modeling strategies to semiparametric variance estimators of a family of efficient, asymptotically linear estimators of causal effects, defined by counterfactual exposure contrasts. Augmenting the inferential stability of these estimators in high-dimensional settings yields a data adaptive approach for robustly uncovering stable causal associations, even when sample sizes are limited. Our generalized variance estimator is evaluated against appropriate alternatives in numerical experiments, and an open source R/Bioconductor package, biotmle, is introduced. The proposal is demonstrated in an analysis of high-dimensional DNA methylation data from an observational study on the epigenetic effects of tobacco smoking.

Keywords

Variance shrinkage; semiparametric estimation; nonparametric inference; efficient estimation; causal machine learning; differential expression; differential methylation

Corresponding author: Nima S Hejazi, Department of Biostatistics, T.H. Chan School of Public Health, Harvard University, 655 Huntington Ave., Boston, MA 02115, USA. nhejazi@hsph.harvard.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

1 Introduction

High-dimensional biomarker data is now routinely collected in observational studies and randomized trials in the biomedical and health sciences. The statistical analysis of such data often relies on parametric modeling efforts that allow covariate adjustment to obtain inference in samples that are small or moderately sized relative to biomarker dimensionality. By treating each biomarker as an independent outcome, standard differential expression analyses fit biomarker-specific linear models while adjusting for potential baseline confounders in the model's postulated form, capturing the effect of a common exposure on each biomarker when the parametric form is *correctly specified*. While the underlying asymptotic theory of linear models is robust, these techniques have been adapted for use in small-sample settings through variance moderation (or shrinkage) approaches, which stabilize inference on the relevant parameter of the linear model. Motivated by the high costs of sequencing experiments in the past, such methods improve inferential quality, which can be compromised by variance estimates that are too small on account of sample size limitations. The moderated t-statistic, the most popular among such approaches, uses shrinkage to stabilize standard error estimates across many target parameters[1]; its corresponding implementation in the limma software package for the R programming language[2] has been heavily utilized in studies using microarray and next-generation sequencing data.[3,4] We generalize this variance moderation strategy to a broad class of asymptotically efficient estimators compatible with machine learning, increasing their robustness in settings with a limited number of independent units.

Given a high-dimensional biological dataset, a standard differential expression analysis pipeline proceeds by fitting a common-form linear model individually to each of the many candidate biomarkers, using exposure status as the primary independent variable and adjusting for potential confounders of the exposure–outcome relationship by the addition of main terms to the parametric functional form. When sample sizes are small, the moderated t-statistic may be used to stabilize inference by shrinking the biomarker-specific variance estimates towards a common value across the many candidate biomarkers.[1] This improves inference for each biomarker by preventing individual test statistics from spiking on account of overly low (and poor) variance estimates. Still, multiplicity corrections are also necessary to adjust for the testing of many hypotheses.[5] Within this framework, the estimated coefficient of the exposure is taken as an estimate of the scientific quantity of interest—that is, the causal effect of the exposure on the expression of candidate biomarkers. While it is common practice, such an approach is rarely rooted in available scientific knowledge, requiring unfounded assumptions (e.g. postulating an exact linear form) to be introduced by the analyst. A common pitfall in standard practice is misspecification of this parametric form, which leads to the target estimand being misaligned with the motivating scientific question. Only recently have tools from modern causal inference[6] been recognized as offering rigorous solutions to such issues in observational biomarker studies.[7,8]

A rich literature has developed around the construction of techniques that eschew parametric forms, relying instead on developments in non/semi-parametric inference and machine learning[9,10] to avoid the pitfalls of model misspecification. By targeting nonparametric estimands and performing model fitting via automated, data adaptive

regression techniques,[11,12] such non/semi-parametric procedures exhibit flexibility and robustness unavailable to approaches based on the general linear model. Unfortunately, a common limitation in their application is the mutual incompatibility of machine learning-based strategies, convergence rates required for asymptotic statistical inference, and the limited sample sizes often characteristic of biomarker studies. Since non/semi-parametric estimators generally converge at much larger sample sizes than their parametric counterparts,[10] these approaches can suffer from analogous variance estimation instability in even modestly sized studies and thus stand to benefit from variance moderation at larger sample sizes than parametric techniques.

Our principal contribution is an adaptation of a parametric shrinkage estimator of variance, or variance moderation, to derive stabilized inference for data adaptive estimators of nonparametric estimands. Specifically, through the comparison of four non/semi-parametric variance estimation strategies, we demonstrate that a generalized variance shrinkage approach can improve the stability of efficient, data adaptive estimation procedures in small and modestly sized biomarker studies. We introduce a modified reference distribution for hypothesis testing with moderated test statistics, further strengthening the Type-I error control of our biomarker identification strategy. We emphasize that our proposal need not be a competitor to other marginal variance stabilization strategies formulated for non/semi-parametric efficient estimators; rather, it may be coupled with such methods to further stabilize inference by directly improving the quality of individual, biomarker-specific variance estimates.

Our approach may be applied directly to a wide variety of parameters commonly of interest as long as an *asymptotically linear estimator* of the target parameter exists. Such estimators are characterized by their asymptotic difference from the target parameter admitting a representation as the sum of independent and identically distributed, mean-zero random variables (i.e. the estimator's influence function). Asymptotically linear estimators have been formulated for both parametric estimands and nonparametric estimands defined in conjunction with causal models.[10] While our variance moderation approach may be applied in a vast array of problems, its advantages are particularly noteworthy in high-dimensional settings, when the sampling distributions of complex, non/semi-parametric efficient estimators can be erratic and prone to yielding high false positive rates.

The remainder of the present article is organized as follows. First, we briefly introduce elements of non/semi-parametric theory for locally efficient estimation and variance moderation in the traditional modeling paradigm. Next, we detail the proposed approach, including an illustration of generalizing variance shrinkage to two common non/semi-parametric efficient, doubly robust estimators of the average treatment effect, complete with a robustified, moderated test statistic. The results of interrogating the proposed technique in simulation experiments are then presented, demonstrating relative performance against a popular variance-moderated linear modeling strategy and non/semi-parametric efficient estimators without variance moderation. Having characterized the proposed procedure's properties in numerical studies, we go on to demonstrate the application of our variance-moderated doubly robust estimation procedure to evaluate evidence from an observational study[13] on the epigenetic alterations to DNA methylation biomarkers causally associated

with tobacco smoking. We conclude by summarizing our findings and by identifying potential avenues for future investigation.

2 Preliminaries and background

2.1 Data, notation, and statistical model

We consider data generated by typical cohort sampling, where the data on a single observational unit is denoted by the random variable $O = (W, A, Y)$, where $W \in \mathbb{R}^d$ is a d-dimensional vector of baseline covariates, $A \in \{0, 1\}$ is a binary exposure, and $Y = (Y_b, b = 1, \ldots, B) \in \mathbb{R}^b$ is a b-dimensional vector of outcomes (e.g. candidate biomarker measurements). We assume access to n independent copies of O, using P_0 to denote the distribution of O. Further, we assume a nonparametric statistical model $P_0 \in \mathcal{M}$ composed of all distributions subject to some dominating measure, thereby placing no restrictions on the form of P_0. Letting $q_{0,Y}$ denote the conditional density of Y given (A, W), $g_{0,A} := \mathbb{P}(A = 1 \mid W)$ the conditional probability of A given W, and $q_{0,W}$ the density of W, the density of O, p_0, evaluated on a typical observation o, may be expressed $p_0(o) = q_{0,Y}(y \mid A = a, W = w) g_{0,A}(a \mid W = w) q_{0,W}(w)$.

A nonparametric structural equation model (NPSEM) allows for counterfactual quantities of interest to be described by hypothetical interventions on the data-generating mechanism of O.[6] We assume an NPSEM composed of the following system of equations: $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(A, W, U_Y)$, where f_W, f_A, and f_Y are deterministic functions, and U_W, U_A, and U_Y are exogenous random variables. The NPSEM provides a parameterization of p_0 in terms of the distribution of the endogenous and exogenous random variables modeled by the system of structural equations, implying a model for the distribution of counterfactual random variables generated by specific interventions on the data-generating process. For simplicity, we consider a static intervention, defined by replacing f_A with a value a $a \in \mathcal{A} \equiv \{0, 1\}$, the support of A. Such an intervention generates a counterfactual random variable $Y(a) = (Y_{b(a)}, b: 1, \ldots, B)$, defined as the values the B candidate biomarker outcomes would have taken if the exposure A had been set to level $a \in \mathcal{A}$, possibly contrary to fact.

To proceed, we define the target parameter as a variable importance measure based on the statistical functional corresponding, under standard identification assumptions,[6] to a causal parameter. The target parameter $\Psi(P_0)$ is defined as a function Ψ mapping the true probability distribution $P_0 \in \mathcal{M}$ of O into a target feature of interest. Letting P_n denote the empirical distribution of the observed sample, $O_1, \ldots, O_n$, an estimate ψ_n of the target parameter $\psi_0 \equiv \Psi(P_0)$ may be viewed as a mapping from $\mathcal{M}$ to the parameter space Ψ.[10] By casting the target parameter as a feature of the (unobserved) true probability distribution P_0, this definition allows a much richer class of target features of interest than the more restrictive view of considering only coefficients in possibly misspecified parametric forms. Throughout, we use the naught subscript to refer to features of the distribution P_0 (e.g. expectation $\mathbb{E}_0$ w.r.t. P_0, variance $\mathbb{V}_0$ w.r.t. P_0) and the n subscript to refer to features dependent on P_n. For clarity of notation and exposition, we focus on cases where

$O_1, \ldots, O_n$ are i.i.d., noting that the proposed methodology generalizes, with only very minor modifications, to cases in which the observed units are clustered, such as when repeated samples on the same biological unit (i.e. technical replicates) are available.

2.2 Targeted variable importance measures

In the high-dimensional settings common in biomarker discovery studies, the tools of causal inference and non/semiparametric theory may be leveraged to develop efficient estimators of the effect of an exposure on an outcome while flexibly controlling for unwanted effects attributable to potential confounders. Commonly, variable importance analyses seek to derive rankings of the relative importance of candidate biomarkers based on their independent associations with another variable of interest, such as exposure to an environmental toxin or disease status.[9,14,10] Related prior proposals[9] defined a variable importance measure based on the average treatment effect (ATE), a constrast of counterfactual means, as

$$\psi_{0,b} \equiv \Psi_b(P_0) := \mathbb{E}_0[\mathbb{E}_0(Y_b \mid A = 1, W) - \mathbb{E}_0(\mathrm{Y_b} \mid \mathrm{A} = 0, \mathrm{W})] \tag{1}$$

for a given biomarker b. The target parameter of equation (1) is the statistical functional corresponding to the ATE under the identification assumptions standard in causal inference, including no unmeasured confounding and positivity.[6,15] When these assumptions hold, $\psi_{0,b}$ may be interpreted as the causal difference in the mean expression of the biomarker under two counterfactual contrasts defined by static interventions on the binary exposure A[6]; however, even when these assumptions are unsatisfied, the statistical target parameter is endowed with a straightforward interpretation: it is the adjusted mean difference in candidate biomarker expression across exposure contrasts, marginalized over strata of baseline confounders.[10] Throughout, we use the ATE to contextualize our developments, as this parameter is ubiquitous in causal inference (and, accordingly, familiar to many); however, no aspect of our proposal is tied to the ATE.

In efficiency theory, the efficient influence function (EIF) occupies a prominent role, for its asymptotic variance is the non/semi-parametric efficiency bound (i.e. variance lower bound) for all regular and asymptotically linear estimators. Thus, efficient estimators may be constructed as solutions to an estimating equation based on the EIF $D_0(O_i)$. For the biomarker-specific ATE $\psi_{0,b}$, the form of the EIF[9,10] is

$$D_{0,b}(O_i) = \left[\frac{2A_i - 1}{g_0(A_i \mid W_i)}\right]\left(Y_{b,i} - \bar{Q}_{0,b}(A_i, W_i)\right) + \bar{Q}_{0,b}(1, W_i) - \bar{Q}_{0,b}(0, W_i) - \psi_{0,b} \tag{2}$$

In equation (2), $D_{0,b}(O_i)$ is the EIF evaluated at a single observed data unit O_i, $\bar{Q}_{0,b}(A, W) = \mathbb{E}_0(Y_b \mid A, W)$ is the true outcome regression at P_0 (with corresponding estimator $\bar{Q}_{n,b}$) evaluated at values of the intervention $A \in \{0, 1\}$, and $g_0(A \mid W) = \mathbb{P}_0(A = 1 \mid W)$ is the true propensity score at P_0 (with corresponding estimator g_n). Classical estimators[15] of the

ATE (based on, e.g. G-computation or inverse probability weighting) require access to either the propensity score or outcome regression, while non/semi-parametric efficient estimators based on the EIF generally require estimation of both nuisance parameters.

2.3 Data adaptive efficient estimation

Several approaches exist for constructing efficient estimators based on the EIF. Among these, two popular frameworks incorporate data adaptive regression: one-step estimation[16] and targeted minimum loss (TML) estimation.[11,10] Both strategies begin by first estimating the nuisance parameters $(g_0, \bar{Q}_{0,b})$, proceeding to then employ distinct bias-correcting procedures in their second stages. As a property of the ATE's EIF, the resultant estimators, regardless of the framework used, are consistent when either of the nuisance parameters is correctly estimated (i.e. doubly robust) and asymptotically achieve the non/semi-parametric efficiency bound when both are well-estimated in a rate-convergence sense.[16,10,17]

2.3.1 Constructing initial estimators—Both classes of efficient estimators accommodate flexible, data adaptive regression (i.e. machine learning) for the construction of initial estimates of the nuisance parameters $(g_0, \bar{Q}_{0,b})$, sharply curbing the risk for model misspecification. Considering the vast and constantly growing array of machine learning algorithms in circulation, it can be challenging to select a single algorithm or family of learning algorithms for optimal estimation of $(g_n, \bar{Q}_{n,b})$. Two strategies for addressing this challenge include model selection through a combination of cross-validation and loss-based estimation[18,19] and model ensembling.[20] The Super Learner algorithm[12] unifies these strategies by leveraging the asymptotic optimality of cross-validated loss-based estimation[19] to either select a single algorithm or produce a weighted ensemble from a user-specified candidate library via empirical risk minimization of an appropriate loss function. The result is an asymptotically optimal procedure for estimation of the nuisance parameters $(g_n, \bar{Q}_{n,b})$, more aptly capturing their potentially complex functional forms. A modern implementation of the Super Learner algorithm is available in the sl3[21] *R* package.

2.3.2 Efficient estimation—In one-step estimation, the empirical mean of the estimated EIF is added to the initial plug-in estimator, that is, $\psi^{+}_{n,b} = n^{-1}\sum_{i=1}^{n}\bar{Q}_{n,b}(1, W_i) - \bar{Q}_{n,b}(0, W_i) + D_{n,b}(O_i)$, where $D_{n,b}(O_i) = [(2A_i - 1)/g_n(A_i \mid W_i)]\left(Y_{b,i} - \bar{Q}_{n,b}(A_i, W_i)\right) + \bar{Q}_{n,b}(1, W_i) - \bar{Q}_{n,b}(0, W_i) - \psi_{n,b}$ is the EIF evaluated at the initial nuisance parameter estimates $(g_n, \bar{Q}_{n,b})$ and $\psi_{n,b}$ is the substitution estimator implied by G-computation.[15] TML estimation takes the alternative approach of tilting the nuisance parameters of the plug-in estimator to solve critical score equations based on the form of the EIF. The TML estimator is $\psi^{\star}_{n,b} = n^{-1}\sum_{i=1}^{n}\bar{Q}^{\star}_{n,b}(1, W_i) - \bar{Q}^{\star}_{n,b}(0, W_i)$, where $\bar{Q}^{\star}_{n,b}$ is an updated version of the initial estimate $\bar{Q}_{n,b}$ of the outcome regression. The updating procedure perturbs the initial estimate $\bar{Q}_{n,b}$ via a carefully constructed one-dimensional parametric fluctuation model, that is, $logit\left(\bar{Q}^{\star}_{n,b}(A, W)\right) = logit\left(\bar{Q}_{n,b}(A, W)\right) + \epsilon_n h(A, W)$, in which the initial outcome estimate $\bar{Q}_{n,b}(A, W)$ is treated as an offset (i.e. coefficient fixed to 1) and ϵ_n is the coefficient of the auxiliary covariate $h(A, W) = (2A - 1)/g_n(A \mid W)$, which incorporates inverse probability

weights based on the initial propensity score estimate $g_n(A \mid W)$. When g_n takes extreme values (close to the boundaries of the unit interval), the fluctuation model may instead include $h(A, W)$ as a weight, which could improve estimation stability. The TML estimator $\psi^{\star}_{n,b}$ of ψ_b is derived by plugging in the updated estimates $\overline{Q}^{\star}_{n,b}$ into the substitution formula based on G-computation. Owing to their distinct bias-correcting steps, both the one-step estimator $\psi^{+}_{n,b}$ and the TML estimator $\psi^{\star}_{n,b}$ share an asymptotically normal limit distribution

$$\sqrt{n}(\psi_{n,b} - \psi_{0,b}) \underset{D}{\rightarrow} \mathrm{N}(0, \mathbb{V}_0(D_{0,b}(O))) \tag{3}$$

where $\psi_{n,b}$ denotes either $\psi^{\star}_{n,b}$ or $\psi^{+}_{n,b}$. Using this limit distribution, inference based on Wald-style confidence intervals and compatible hypothesis tests is readily attainable.

2.3.3 Variance estimation based on the efficient influence function—

Given their normal limiting distribution (3), a standard variance estimator for asymptotically linear estimators may be formulated based on the scaled empirical variance of the estimated EIF. This variance estimator, $\sigma^2_{n,b} := \mathbb{V}_n D_{n,b} = n^{-1}\sum_{I=1}^{n} D^2_{n,b}(O_i)$, uses the initial estimates of the nuisance parameters, and is a valid, occasionally conservative variance estimator for both the one-step and TML estimators. A popular alternative approach instead uses the empirical variance estimator based on a cross-validated estimate of the EIF, which can address issues of overfitting of nuisance function estimators while removing theoretical entropy conditions for asymptotic inference,[22,23] both of which have contributed to the current prevalence of sample-splitting (i.e. cross-validation, “cross-fitting”). Though this approach improves marginal variance estimates $\sigma^2_{n,b}$, it fails to take advantage of the benefits that pooled variance estimation provides in high-dimensional settings.

Since we advocate for the use of data adaptive regression techniques for nuisance parameter estimation, we wish to draw particular attention to the cross-validated variance estimator based on the EIF. Analogous to the full-sample variance estimator, this estimator is based on the empirical variance of the EIF evaluated at cross-validated initial estimates of the nuisance functions. To define such an estimator, consider the use of K-fold cross-validation, denoting by $V_1, \ldots, V_K$ a random partition of the index set $\{1, \ldots, n\}$ into K validation sets of roughly the same size. That is, $V_k \subset \{1, \ldots, n\}$, $\cup_{k=1}^{K} V_k = \{1, \ldots, n\}$, and $V_k \cap V_{k'} = \varnothing$ for $k \neq k'$. For each k, its training sample is $\mathcal{T}_k = \{1, \ldots, n\} \setminus V_k$. Let $(g_{n,k}, \overline{Q}_{n,k,b})$ be the estimators of $(g_0, \overline{Q}_{0,b})$ constructed by fitting a data adaptive regression procedure using only data available in the training sample $\mathcal{T}_k$. Then, letting $j(i)$ denote the index of the validation set containing observation i, the empirical variance of the cross-validated EIF is $\sigma^2_{n,\mathrm{CV},b} = \mathbb{V}_n D_{n,j(i),b}$, where $D_{n,j(i),b}$ is the EIF evaluated at $(g_{n,j(i)}, \overline{Q}_{n,ji(i),b})$. We explore the use of this variance estimator and any advantages it may confer in our context in the sequel.

2.4 Parametric variance moderation

Variance moderation has been established as a promising and useful tool for stabilizing test statistics. The general methodology consists in the application of a shrinkage estimator to

the individual variance estimates across a large number of (related) hypothesis tests. The moderated t- and F-statistics[1] are perhaps the most commonly used examples of variance moderation approaches in differential expression analysis, though their original formulation is tied to the general linear model. In that context, a typical differential expression analysis would fit B linear models $\hat{Y}_b = \hat{\beta}_{0,b} + \hat{\beta}_{1,b}A + \hat{\beta}_{2,b}W$, using a standard or moderated test statistic to assess the association of A on each of the B biomarkers independently. This moderated t-statistic[1] is

$$\tilde{t}_b = \frac{\hat{\beta}_{1,b}}{\tilde{\sigma}_b} \text{ where } \tilde{\sigma}_b^2 = \frac{d_0\hat{\sigma}_0^2 + d_b\hat{\sigma}_b^2}{d_0 + d_b} \tag{4}$$

in which d_b and d_0 are the degrees of freedom for the b^{th} biomarker and the remaining $(B-1)$ biomarkers, respectively, and $\hat{\sigma}_b$ is the standard deviation for the b^{th} biomarker while $\hat{\sigma}_0$ is the standard deviation across all of the other $(B-1)$ biomarkers.

The resultant test statistic has much the same interpretation as an ordinary t-statistic, though its standard error is now shrunken towards a common value (i.e. moderated) across all biomarkers. This form of moderation prevents the test statistic from spiking even when the variance estimate for the biomarker in question, $\hat{\sigma}_b^2$, is too small. Accordingly, $\tilde{t}_b$ is said to be more stable than its non-stabilized analog t_b. The process of generating p-values for the moderated t-statistic is analogous to that of the ordinary t-statistic, with the only difference being that the degrees of freedom may be inflated to offset the increased robustness induced by moderation.[1] The approach was introduced in the limmaR package, available via the Bioconductor project[3,24]; it remains extremely popular for biomarker identification and differential expression analysis. Next, we adapt this approach for use with the efficient estimators previously described.

3 Semi-parametric variance moderation

Application of TML estimation to construct targeted variable importance estimates for a given set of biomarkers has been previously considered[9]; however, marginal estimates of variable importance are often insufficient or unreliable for deriving joint inference in high-dimensional settings. Such approaches suffer significantly from the instability of standard error estimates in settings with limited sample sizes, erroneously identifying differentially expressed biomarkers. This considerably limits their utility in high-dimensional biomarker studies. In order to obtain stable joint inference on a targeted variable importance measure across many biomarkers $b = 1, \ldots, B$, we propose the use of variance moderation, which may be achieved by applying the moderated t-statistic[1] to shrink biomarker-specific estimates of sampling variability based on the estimated EIF towards a stabilized, pooled estimate across biomarkers.

As inference for ψ_b is based on individual variability estimates $\sigma_{n,b}$ (each derived from the corresponding biomarker-specific EIF $D_{0,b}$), our proposed generalization applies moderation to the estimated EIF $D_{n,b}$, yielding a *moderated EIF* $\tilde{D}_{n,b}$. The resultant moderated variance

estimate $\tilde{\sigma}^2_{n,b}$ is then the empirical variance of $\widetilde{D}_{n,b}$. The resultant stabilized variability estimates $\tilde{\sigma}_{n,b}$ may directly be used in the construction of Wald-style confidence intervals or the evaluation of hypothesis tests. Consider B independent tests with null and alternative hypotheses $H_0: \psi_{0,b} = 0$ and $H_1: \psi_{0,b} \neq 0$, and let $\psi_{n,b}$ denote either the one-step or TML estimator of $\psi_{0,b}$; then, our proposal is as follows:

1. Optionally, reduce the set of hypotheses by a filtering procedure, which may reduce the computational burden imposed by using flexible regression strategies for nuisance parameter estimation across many biomarker outcomes. As long as this initial filtering procedure does not affect the candidate biomarker rankings, its effect may be readily accounted for in post-hoc multiple hypothesis testing corrections.[25]

2. For each biomarker, generate non/semi-parametric efficient estimates $\psi_{n,b}$ of $\psi_{0,b}$ and corresponding estimates of the EIF $D_{n,b}(O_i)$, evaluated at the initial estimates of the nuisance parameters $(g_n, \overline{Q}_{n,b})$.

3. Apply variance moderation across the biomarker-specific EIF estimates $(D_{n,b}: b = 1, \ldots, B)$ (e.g. via the limma R package[3]), constructing moderated EIF estimates $\tilde{\sigma}^2_{n,b}$ for each biomarker. The moderated variance estimates are constructed by shrinking each $\sigma^2_{n,b}$ towards the group variance across all other $(B - 1)$ biomarkers as per (4). Note that the variance moderation step is asymptotically inconsequential, that is, $\tilde{\sigma}_{n,b} \to \sigma_{n,b}$ as $n \to \infty$, and the limit distribution (3) holds.

4. For each biomarker-specific estimate $\psi_{n,b}$ of the target parameter $\psi_{0,b}$, construct a moderated t-statistic $(\tilde{t}_b: b = 1, \ldots, B)$ based on the corresponding moderated standard error estimate $\tilde{\sigma}_{n,b}$. The test statistic $\tilde{t}_b = \psi_{n,b}/\tilde{\sigma}_{n,b}$ may be used to evaluate evidence for the null hypothesis $H_0: \psi_{0,b} = 0$ of no treatment effect against the alternative $H_1: \psi_{0,b} \neq 0$. While the t-distribution with adjusted degrees of freedom was the originally proposed reference distribution[1] for such a test statistic, we advocate instead for the use of a standardized logistic distribution (with zero mean and unit variance). This alternative reference distribution exhibits subexponential tail behavior, allowing for improved conservative inference. This is useful in high-dimensional settings, where the joint distribution of all $(\tilde{t}_b: b = 1, \ldots, B)$ test statistics may fail to converge quickly enough in n to a B-dimensional multivariate normal or t-distribution, thus thwarting attempts to appropriately control the joint error. By contrast, the heavier, subexponential tails of the logistic distribution provide more robust error control. Alternative approaches to conservative inference based on, for example, concentration inequalities[26] or Edgeworth expansions,[27] may also be suitable and are compatible with our proposal.

5. Use a multiple testing correction to obtain accurate simultaneous inference across all B biomarkers. A common approach is to use the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR),[28] which controls Type-I

error proportion in expectation in high-dimensional settings under conditions commonly considered acceptable in computational biology applications. This is not the only available choice and our proposal readily accommodates alternatives.

Intuitively, our proposed variance moderation procedure shrinks aberrant variability estimates towards the center of their joint distribution, with a particularly noticeable reduction of Type-I error when the sample size is small and variance estimates unstable. Practically, this approach limits the number of statistically significant findings driven by unstable estimates of the variance of $\psi_{n,b}$, for the exact same reason discussed earlier.

What's more, our proposal is convenient on account of its straightforward application to existing variance estimators based on the EIF and valid in all cases where asymptotically linear estimators may be constructed. Since our proposed procedure consists in a moderated variance estimator based on the empirical variance of the estimated EIF, we stress that providing enhanced Type-I error rate control is only guaranteed for multiple testing procedures that are based on marginal hypothesis tests, as opposed to alternative techniques (e.g. permutation and resampling methods) that directly target the joint distribution of test statistics.[5] To enhance accessibility, we have made available an open source software implementation, the biotmle package,[29,30] available for the R language and environment for statistical computing[2] through the Bioconductor project[24] for computational biology and bioinformatics.

4 Simulation studies

We evaluated our variance moderation strategy based on its Type-I error control as assessed by the FDR.[28] We focus on the FDR owing to its pervasive use in addressing multiple hypothesis testing in high-dimensional biology; however, our approach is equally compatible with most post-hoc multiple testing corrections (e.g. Bonferroni's method to control the family-wise error rate). We assessed the relative performance of several data adaptive non/semi-parametric estimators of the ATE, each using identical point estimation methodology but different marginal variance estimators, and a single linear modeling strategy in terms of their accuracy for joint inference. We considered the performance of five variance estimation strategies (colors and shapes reference the figures appearing later): (1) "standard" variance moderation using $\tilde{\sigma}_b^2$ via the limma R package[3] with a main terms linear model (red circle); (2) a TML estimator $\psi_{n,b}^{\star}$ coupled with the empirical variance of the full-sample EIF $\sigma_{n,b}^2$ (yellow triangle); (3) a TML estimator $\psi_{n,b}^{\star}$ coupled with the empirical variance of the cross-validated EIF $\sigma_{n,\mathrm{CV},b}^2$ (cyan triangle); (4) a TML estimator $\psi_{n,b}^{\star}$ using our variance moderation of the full-sample EIF $\tilde{\sigma}_{n,b}^2$ (yellow square); and (5) a TML estimator $\psi_{n,b}^{\star}$ using our variance moderation of the cross-validated EIF $\tilde{\sigma}_{n,\mathrm{CV},b}^2$ (cyan square). For the cross-validated variance estimators, we chose two-fold cross-validation based on a conjecture that larger validation fold sizes would yield more conservative variance estimates. We note that the one-step and TML estimators are asymptotically equivalent and share a variance estimator, yet we use the TML estimator on account of there being some evidence of enhanced finite-sample performance.[10] The TML estimators and their corresponding variance estimators were based on the implementations in the drtmle[31,32] and biotmle[29,30] R packages. To isolate the effect

of variance moderation on FDR control, all efficient estimator variants used the logistic reference distribution.

For these experiments, we simulated data from the following data-generating mechanism. First, two baseline covariates are independently drawn as $W_1 \sim Uniform(0, 1)$ and $W_2 \sim Uniform(0, 1)$. Next, the exposure A is drawn, conditionally on $\{W_1, W_2\}$, from $A \mid W \sim Bernoulli[expit(0.5 + 2.5W_1 - 3W_2)]$. Finally, biomarker expression Y_b is generated, conditionally on $\{A, W_1, W_2\}$, by either $Y_{\text{null}} \mid A, W = 2 + W_1 + 0.5W_2 + W_1 \cdot W_2 + \epsilon_1$ or $Y_{\text{strong}} \mid A, W = 2 + W_1 + 0.5W_2 + W_1 \cdot W_2 + 5A + \epsilon_2$. Throughout, $expit(x) = \{1 + \exp(-x)\}^{-1}$, $\epsilon_1 \sim Normal(0, 1)$, and $\epsilon_2 \sim Normal(0, 0.2)$. The data on a single observational unit are denoted by the random variable $O = (W_1, W_2, A, (Y_b : 1, \cdots, B))$, where each biomarker $(Y_b : 1, \cdots, B)$ is generated from Y_{strong} or Y_{null} depending on the setting. Note the shared functional form of the outcome models, in particular that the interaction term between $\{W_1, W_2\}$ gives rise to model misspecification issues when linear regression is employed out-of-the-box. This design choice draws attention to the advantages of relying upon non/semi-parametric efficient estimation frameworks capable of incorporating data adaptive regression strategies (i.e. machine learning) in nuisance estimation.

For applications in which the exposure mechanism exhibits a lack of natural experimentation (i.e. positivity violations), estimation of the exposure mechanism $g_n(A \mid W)$ can yield values extremely close to the boundaries of the unit interval. Such extreme estimates compromise the performance of data adaptive non/semi-parametric estimators,[33] in part due to the instability of estimated inverse probability weights. Often, practical violations of the positivity assumption occur when the exposure A is strongly related to the baseline covariates W, which manifests as an apparent lack of experimentation of the exposure across covariate strata. To assess the impact of such violations on variance estimation, we replace the exposure mechanism with $A \mid W \sim Bernoulli(expit(0.5 + 2.5W_1 - 3W_2 - 2))$ in a few scenarios. Unlike the exposure mechanism above, which allows a minimum exposure probability of 0.076, this exposure mechanism allows a minimum exposure probability of 0.011, leading to positivity issues that may exacerbate bias and variance instability in high dimensions.

To ensure compatibility of each of the efficient estimator variants, initial estimates of the nuisance functions $g_n(A \mid W)$ and $\bar{Q}_{n,b}(A, W)$ were constructed using the Super Learner[12] algorithm. The SuperLearner R package[34] was used to construct ensemble models from a library of candidate algorithms that included linear or logistic regression, regression with Bayesian priors, generalized additive models,[35] multivariate adaptive regression splines,[36] extreme gradient boosted trees,[37] and random forests.[38]

Here, we consider settings in which the exposure affects 10% or 30% of all biomarkers. In each scenario, $B = 150$ biomarkers are drawn from the equations for Y_{null} and Y_{strong} in differing proportions. In any given simulation, we consider observing n i.i.d. copies of O for one of four sample sizes $n \in \{50, 100, 200, 400\}$. Overall, we consider scenarios in which the number of biomarkers exceeds the sample size as well as settings outside the high-dimensional regime, that is, $n/p = \{1/3, 2/3, 4/3, 8/3\}$. The former set of scenarios

emphasizes the utility of variance moderation when $p > n$, while the latter demonstrates its negligible effect in larger samples.

Results are reported based on aggregation across 300 Monte Carlo repetitions for each scenario. In aggregate, these scenarios are used to evaluate the degree to which each of the five variance estimation strategies controls the FDR. Throughout, we restrict our attention to control of the FDR at the 5% level, as this is most commonly used in practice and the choice of threshold has no impact on our proposed procedure. A few additional scenarios are considered in the Supplemental Materials, including the relative estimator performance in cases with no exposure effect and when there is a weaker exposure effect than in the presently considered setting.

We begin with a scenario in which the effect of the exposure on biomarker expression is strong, when the effect is either relatively rare (10% of biomarkers) or fairly common (30% of biomarkers). In the rare effect setting, expression values for the affected 10% of biomarkers are generated by Y_{strong} while the values for the remaining 90% arise from Y_{null}. Here, we expect the efficient estimators with EIF-based variance estimation strategies (whether moderated or not) to exhibit FDR control approaching the nominal rate with increasing sample size while reliably recovering truly differentially expressed biomarkers. Due to bias arising from misspecification of the outcome model, the moderated linear model is expected to perform poorly. The performance of the estimator variants is presented in Figure 1.

As expected, variance-moderated hypothesis tests based on linear modeling fail to control the FDR at the 5% rate due primarily to model misspecification. The efficient estimators based on the EIF exhibit reasonable performance, with the full-sample variance estimators achieving the nominal rate by $n = 400$ and the cross-validated variants consistently controlling the FDR more stringently than the nominal rate. Examination of the false discovery proportions reveals that variance moderation provides some benefit in improving FDR control at $n = 50$, though this disappears quickly with increasing sample size. While the true positive rates indicate good performance of all candidate procedures (though the cross-validated variants are less reliable at smaller sample sizes), the true negative rates demonstrate the consistent performance of the cross-validated variants, performance improving with sample size for the full-sample estimators, and degrading performance for the linear model.

We now turn to a setting in which the exposure mechanism is prone to positivity violations. In this case, the full-sample EIF-based variance estimators are expected to exhibit relatively poor performance due to estimation instability in the inverse probability weights; however, the cross-validated variants are expected to provide FDR control at the nominal rate without sacrificing power. Figure 2 presents the estimator performance.

As before, linear model-based hypothesis testing fails to control the FDR at the 5% rate (owing to model misspecification). Positivity violations in the exposure mechanism result in the full-sample EIF-based estimators yielding poor FDR control as well. Their cross-validated counterparts fare significantly better, achieving control at the nominal rate

by $n = 200$. Both the FDR and false discovery proportion panels illustrate that variance moderation of the efficient estimators modestly but *uniformly* improves their FDR control, regardless of the use of sample-splitting in nuisance estimation. Consideration of the true positive rates reveals good performance of all candidate procedures (again, the cross-validated variants are slightly over-conservative). The true negative rates show very strong control from the cross-validated variants and worse but improving performance from the full-sample estimators; the linear model displays unreliable, degrading performance. The protective effect of variance moderation is made clear by the true negative rates.

Next, we turn to a setting in which the exposure has a strong effect on a larger proportion of biomarkers. This scenario is constructed by generating expression values for 30% of biomarkers from Y_{strong} and the remaining 70% from Y_{null}. We begin with the exposure mechanism not prone to positivity violations, in which case both the full-sample and cross-validated efficient estimators are expected to exhibit FDR control near the nominal rate, regardless of variance moderation. Due to model misspecification, the moderated linear model is expected to exhibit poor FDR control. Figure 3 visualizes the performance of the candidate procedures.

Given that the exposure effect on biomarkers is more common, all of the estimator variants fare comparatively better than in the rarer effect scenario considered previously. As before, the poor performance of the linear modeling strategy is caused by model misspecification bias. In comparison, the efficient estimators all exhibit better performance, with the full-sample variance estimators controlling the FDR at nearly the nominal rate and the cross-validated variants providing more stringent control. As with the prior setting summarized in Figure 1, the effect of variance moderation on FDR control is subtle, though examination of the lower panel of Figure 3 reveals the stronger error rate control that variance moderation achieves. While the true positive rates reveal good performance from all candidate estimators by $n = 100$, the true negative rates show slightly better control from the cross-validated variants (relative to their full-sample counterparts); the linear model shows poor performance at $n = 50$ and only degrades considerably thereafter.

Finally, we again consider an analogous setting in which the exposure mechanism has positivity issues. As before, the linear modeling procedure is expected to perform poorly. The efficient estimators with full-sample EIF-based variance estimation ought to perform relatively poorly due to estimation instability (from positivity violations) while the cross-validated variants are expected to provide close-to-nominal FDR control. Figure 4 presents the results of examining the estimator variants in this setting.

The upper panel of Figure 4 corroborates our expectations about the linear modeling strategy's potential to yield erroneous discoveries. While the linear model outperforms a subset of the efficient estimators at $n = 50$, its performance degrades sharply thereafter. The efficient estimators using full-sample EIF-based variance estimation display relatively poor control of the FDR, failing to achieve the nominal rate but maintaining their performance across sample sizes (unlike the linear model). The estimator variants using cross-validated EIF-based variance estimation exhibit far improved control of the FDR, nearly achieving the nominal rate in smaller sample sizes and controlling the FDR more stringently in larger

samples. A quick examination of the lower panel of the figure makes clear the modest improvements to error rate control that variance moderation provides. In particular, the true positive rates are quite reliable for all candidate estimators, though the cross-validated estimator variants are somewhat over-conservative in smaller samples. By comparison, the true negative rates reveal the stronger control that variance moderation confers for both the cross-validated and full-sample estimator variants, and highlights the predictably poor performance of the linear modeling strategy. Echoing results of the experiments presented in Figure 2, variance moderation improves FDR control irrespective of whether sample-splitting is used.

Additional simulation experiments and their results are presented in the Supplemental Materials. There, we consider two distinct scenarios, one in which there is no effect of exposure at all (i.e. the “global null”) and another in which the effect of the exposure is attenuated relative to the scenario considered here. In the former setting, we find that the cross-validated variance estimators provide error rate control in line with the nominal rate of 5% (though, at times, they are conservative) while the full-sample analogs break down under positivity violations. In the latter scenario, the cross-validated estimators are uniformly conservative while their full-sample counterparts provide control at or very near the nominal rate, regardless of positivity violations. Overall, variance moderation improves error rate control uniformly across these scenarios too, just as it does in the results discussed above. Altogether, our numerical investigations demonstrate the advantages conferred by applying variance moderation to non/semi-parametric efficient estimators in settings with limited sample sizes and a relatively large number of outcomes. In our experiments, the efficient estimators have access to an eclectic library of machine learning algorithms for nuisance estimation, significantly reducing the risk of model misspecification bias. Generally, the full-sample EIF-based variance estimators exhibit poorer FDR control than their cross-validated counterparts, suggesting a stabilizing effect of sample-splitting on variance estimation, which itself pairs with variance moderation. Our results reveal that variance moderation can have substantial benefits in settings with positivity issues, which occur often in observational studies. Overall, our findings suggest that variance moderation can prove a useful and, at times, powerful tool for modestly improving FDR control in high-dimensional settings, without adversely affecting the recovery of truly differentially expressed biomarkers, and is especially useful in high-dimensional settings when paired with cross-validation.

5 Application in an observational smoking exposure study

We now apply our variance-moderated efficient estimation strategy to examine evidence for differential methylation of CpG sites in whole blood as a result of voluntary smoking exposure. Data for this illustrative application come from an observational exposure study that enrolled 253 healthy volunteer participants between 1993 and 1995 from the general population in Chapel Hill and Durham, North Carolina. Among these participants, 172 self-reported as smokers and 81 as nonsmokers (defined as having smoked fewer than 100 cigarettes in their lifetime). For all participants, a limited set of baseline covariates (a mix of continuous and discrete variables), including biological sex, race/ethnicity (minority status), and age, were recorded. The study protocol and details on processing of biological samples have been previously detailed[39,40,13]; we encourage the interested reader to refer

to these publications for further details. DNA methylation levels of patients' whole blood DNA samples were measured with the Infinium Human Methylation 450K BeadChip (Illumina, Inc.), designed to measure methylation at ≈450,000 CpG sites across the human genome. Prior analytic efforts[13] normalized the raw DNA methylation data via the ChAMP procedure[41,42] and deposited the processed β-values on the NCBI's Gene Expression Omnibus (accession no. GSE85210). In our re-analysis of this study, we used these publicly available DNA methylation data, paired with phenotype data provided by the study team.

For our differential methylation analysis, we used the aforementioned baseline covariates as well as "pack-years" (self-reported packs of cigarettes multiplied by years spent smoking) to adjust for potential baseline confounding of the effect of smoking on DNA methylation. That DNA methylation varies strongly across cell types has been well-studied and documented. Accordingly, we followed standard practice in adjusting for cell-type composition of samples from which DNA was collected by normalization against "gold standard" reference datasets,[43,44] accounting for the relative abundance of CD4+ and CD8+ T-cells, natural killer cells, B-cells, monocytes, and granulocytes. This form of adjustment disentangles the effect of smoking on DNA methylation from the unwanted variation in DNA methylation across cell types from which DNA samples were harvested. Our differential methylation analysis strategy is summarized as follows.

First, the set of roughly 450,000 CpG sites was narrowed down by applying the moderated linear modeling strategy of the limma R package [3] to assess any association of differential methylation with smoking, controlling for baseline covariates in the adjustment set; the 2537 CpG sites with unadjusted p-values below the 5% threshold were advanced to the following stage. Next, using the biotmleR package,[29,30] our variance-moderated non/semi-parametric efficient TML estimator was applied to evaluate evidence for differential methylation attributable to smoking (based on the ATE), again adjusting for the set of potential baseline confounders. Estimation of the nuisance parameters $(g_n, \bar{Q}_{n,b})$ was performed using two-fold cross-validation, and the Super Learner ensemble modeling algorithm[12,34] was used to generate out-of-sample predictions from a library of candidate algorithms that included main-terms GLM regression, multivariate adaptive regression splines,[36] and random forests,[38] among others.

Moderated test statistics were constructed to evaluate the null hypothesis of no ATE at each CpG site, and testing multiplicity was accounted for by adjusting the marginal p-values via Holm's procedure,[45] thereby controlling the family wise error rate (FWER). Marginal p-values for each CpG site were generated by using the standardized normal distribution as reference for the site-specific test statistics (the centered logistic distribution proved too conservative when paired with the FWER metric); moreover, Holm's procedure was chosen over alternative FWER-controlling procedures as its rank-based nature satisfies previously outlined requirements for error rate control in multi-stage analyses.[25] Our choice of FWER prioritizes conservative joint inference, complementing the more lenient reference distribution and highlighting our proposal's flexibility. Our analysis tagged 1173 CpG sites as differentially methylated by voluntary smoking exposure.

The significantly differentially methylated CpG sites are located within the *AHRR, ALPPL2/ALP1, MYO1G, F2RL3*, GFI1, *IER3, HMHB1, ITGAL, LMO7, GPR15, NCOR2, RARA, SPOCK2, HOX* cluster, and RUNX3 genes, among others, agreeing with a prior analysis of these data.[13] Many of these genes have been linked to disease ontology categories like hemotologic cancer, cardiovascular system disease, hematopoietic system disease, and nervous system cancer.[13] In particular, the most significantly differentially methylated CpG site, cg05575921, located in the AHRR gene, has been identified in over 30 epigenome-wide association studies on smoking exposure in both blood and lung tissues.[46]Decreased methylation at this site is widely viewed as a robust biomarker of smoking exposure[46] and is associated with increased lung cancer risk.[47–50]. Table S1 in the Supplemental Materials presents the top 50 differentially methylated CpG sites.

Despite the close agreement between the top set of differentially methylated CpGs revealed by our analysis and those identified in prior analyses, we questioned the stability of our proposal for real-world data analysis. To assess this, we designed and conducted an empirical sensitivity analysis that subsampled study units to capture the effect of data removal on the ranking of differentially methylated CpG sites. The procedure was carried out by sampling without replacement {25%, 50%, 75%} of study units, performing our proposed analysis (as described above) to generate a ranked list of CpG sites, and comparing these top CpG sites against those identified in the complete-data analysis. Since the sensitivity of the preliminary filtering step to subsampling does not relate directly to our procedure's stability, we restricted each of these analyses only to the 2537 CpG sites that passed the filtering step of the complete-data analysis. For each subsampling proportion, this sensitivity analysis strategy was repeated 10 times, allowing for the frequency with which CpGs were tagged as differentially methylated to be evaluated. Figure 5 displays the results of our sensitivity analysis.

Cursory examination of Figure 5 reveals that our findings concerning the top 30 differentially methylated CpG sites are robust to a loss of 25% of study units, as the median adjusted p-values of all of these CpG sites exceed the 5% detection threshold at the 75% subsampling level. Upon further reductions in sample size, the differential methylation signal is still fairly reliable: the median adjusted p-values for ≈75% of the CpG sites (the top 23) exceed the detection threshold even when 50% of study units have been removed. Finally, this form of evidence for differential methylation shows that the top 6 CpG sites identified by our analysis are robust to a loss of as much as 75% of the data, meaning that these same CpGs could have been tagged as differentially methylated had the study included as few as 64 units (instead of the 253 units actually enrolled). Note that while the adjusted p-values reported for each of the 30 CpGs in the figure are the medians across the 10 iterations for each of the subsampling schemes, those for the complete-data analysis are not medians (i.e. that analysis was only run once). Figure S5 in the Supplemental Materials presents an extension of Figure 5, showing how the minimum, median, and maximum adjusted p-values vary across subsampling schemes for the top 30 differentially methylated CpGs. Altogether, this sensitivity analysis demonstrates that our differential methylation procedure reliably recovers evidence for biologically meaningful findings, with power only beginning to degrade significantly with major losses in sample size.

6 Discussion

We have proposed a novel procedure for stabilizing non/semi-parametric efficient estimators of scientifically relevant statistical parameters, combining distinct lines of inquiry on variance moderation and sample-splitting principles in the process. Our variance moderation procedure may be applied directly to the standard variance estimator of regular and asymptotically linear estimators in the nonparametric model, that is, the empirical variance of the estimated efficient influence function. Such asymptotically efficient estimators are capable of incorporating machine learning in nuisance estimation, curbing risks of model misspecification bias, which imposes a significant limitation upon the reliability of parametric modeling approaches. Our variance moderation technique improves the inferential stability of hypothesis testing based on these efficient estimators in high-dimensional settings, and, when combined with cross-validation, it is capable of providing reliably conservative joint inference. Our proposal amounts to a semi-automated procedure for using these state-of-the-art estimators to obtain valid joint inference in high-dimensional biomarker studies while circumventing the pitfalls of model misspecification bias, sampling distribution instability, and anti-conservative variance estimation. Despite its being near-automated, our proposal leaves several key decisions to the data analyst, including the choice of variance estimator (i.e. empirical variance of full-sample versus cross-validated efficient influence function), reference distribution (i.e. multivariate normal or logistic), and multiple testing correction metric (i.e. FDR vs. FWER).The exact choices of these must be made based on the motivating scientific application and the degree to which the analysis in question fulfills exploratory aims.

Our demonstration of this proposal focused on efficient estimators of the average treatment effect; however, the outlined procedure can be readily adapted to any regular and asymptotically linear estimator, accommodating extensions to a wide variety of parameters of scientific interest. Notable areas for future adaptation of this methodology include recently developed estimators of the causal effects of continuous exposures[51,52] and those of causal mediation effects tailored for path analysis.[53,54] Our simulation experiments highlight the benefits conferred by our strategy, both in conjunction with and in the absence of sample-splitting, showing that variance moderation can modestly but uniformly improve Type-I error control in several common scenarios. In a secondary re-analysis of DNA methylation data from an observational study on the epigenetic effects of smoking, we show our procedure to be capable of recovering differentially methylated CpG sites identified in prior analyses and validated in biological experiments; moreover, a sensitivity analysis reveals the findings of our approach to be highly stable even with artificially diminished sample sizes. Given the utility of the procedure, we have developed the free and open source biotmle R package[29,30] and contributed it to the Bioconductor project,[24] making this novel strategy easily accessible to the computational biology scientific community.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank MT Smith, N Rothman, and Q Lan for helpful discussions about an alternative real-world data analysis example. We thank D Bell for providing the data used in the real-world data analysis and for helpful correspondence on the study details for the application presented. We are grateful to S Dudoit for numerous helpful discussions regarding data visualization, sensitivity analysis, and the presentation of results.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: NSH was supported in part by the National Institute of Environmental Health Sciences [award no. R01-ES021369] and the National Science Foundation [award no. DMS 2102840]. PB was funded by the Fonds de recherche du Québec - Nature et technologies, and the Natural Sciences and Engineering Research Council of Canada. MJvdL was partially supported by a grant from the National Institute of Allergy and Infectious Diseases [award no. R01 AI074345]. MT Smith, N Rothman, and Q Lan also received support from the National Institute of Environmental Health Sciences [award no. P42-ES004705] and the National Cancer Institute.

References

1. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3: 1–25.
2. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2022. https://www.R-project.org/.
3. Smyth GK. Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, 2005. pp. 397–420.
4. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 2014; 15: R29. [PubMed: 24485249]
5. Dudoit S and van der Laan MJ. Multiple testing procedures with applications to genomics. New York, NY: Springer, 2008.
6. Pearl J Causality: models, reasoning, and inference. Cambridge University Press, 2000.
7. Reifeis SA, Hudgens MG, Civelek M, et al. Assessing exposure effects on gene expression. Genet Epidemiol 2020; 44: 601–610. [PubMed: 32511796]
8. Reifeis SA. Causal Inference for Observational Genomics Data. PhD Thesis, University of North Carolina at Chapel Hill, 2020.
9. Bembom O, Petersen ML, Rhee SY, et al. Biomarker discovery using targeted maximum-likelihood estimation: application to thetreatment of antiretroviral-resistant HIV infection. Stat Med 2009; 28: 152–172. [PubMed: 18825650]
10. van der Laan MJ and Rose S. Targeted learning: Causal inference for observational and experimental data. New York, NY: Springer Science & Business Media, 2011.
11. van der Laan MJ and Rubin D. Targeted maximum likelihood learning. Int J Biostat 2006; 2.
12. van der Laan MJ, Polley EC and Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007; 6.
13. Su D, Wang X, Campbell MR, et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. PLoS ONE 2016; 11.
14. Tuglus C and van der Laan MJ. Targeted methods for biomarker discovery. In: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, 2011. pp. 367–382.
15. Hernán MA and Robins JM. Causal inference: what if. Boca Raton, FL: CRC Press, 2022.
16. Bickel PJ, Klaassen CA, Ritov Y, et al. Efficient and adaptive estimation for Semiparametric models. Baltimore, MD: Johns Hopkins University Press, 1993.
17. Kennedy EH.: Semiparametric theory and empirical processes in causal inference. In Statistical Causal Inferences and Their Applications in Public Health Research. Springer, 2016. pp. 141–167.
18. van der Laan MJ, Dudoit S and Keles S. Asymptotic optimality of likelihood-based cross-validation. Stat Appl Genet Mol Biol 2004; 3: 1–23.
19. Dudoit S and van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Stat Methodol 2005; 2: 131–154.
20. Breiman L Stacked regressions. Mach Learn 1996; 24: 49–64.

21. Coyle JR, Hejazi NS, Malenica I, et al. sl3: modern pipelines for machine learning and super learning, 2022. doi:10.5281/zenodo.1342293. https://github.com/tlverse/sl3. R package version 1.4.4.
22. Klaassen CA. Consistent estimation of the influence function of locally asymptotically linear estimators. Ann Stat 1987; 0: 1548–1562.
23. Zheng W and van der Laan MJ. Cross-validated targeted minimum-loss-based estimation. In: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, 2011. pp. 459–474.
24. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5: 1–6.
25. Tuglus C and van der Laan MJ. Modified FDR controlling procedure for multi-stage analyses. Stat Appl Genet Mol Biol 2009; 8.
26. Boucheron S, Lugosi G and Massart P. Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press, 2013.
27. Gerlovina I, van der Laan MJ and Hubbard AE. Big data, small sample: Edgeworth expansions provide a cautionary tale. Int J Biostat 2017; 13.
28. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Statistical Methodology) 1995; 57: 289–300.
29. Hejazi NS, Cai W and Hubbard AE. biotmle: targeted learning for biomarker discovery. J Open Sour Softw 2017; 2.
30. Hejazi NS, van der Laan MJ and Hubbard AE. biotmle: targeted learning with moderated statistics for biomarker discovery, 2020. doi:10.18129/B9.bioc.biotmle. https://bioconductor.org/packages/biotmle. R package version 1.12.0.
31. Benkeser DC and Hejazi NS. drtmle: doubly-robust nonparametric estimation and inference, 2022. doi:10.5281/zenodo.844836. R package version 1.1.1.
32. Benkeser D and Hejazi NS. Doubly-robust inference in R using drtmle. Under review at Observational Studies, 2022.
33. Moore KL, Neugebauer R, van der Laan MJ, et al. Causal inference in epidemiological studies with strong confounding. Stat Med 2012; 31: 1380–1404. [PubMed: 22362629]
34. Polley EC, LeDell E, Kennedy CJ, et al. SuperLearner: super learner prediction, 2019. https://github.com/ecpolley/SuperLearner. R package version 2.0-26-9000.
35. Hastie TJ and Tibshirani RJ. Generalized additive models. Routledge, 1990.
36. Friedman JH, et al. Multivariate adaptive regression splines. Ann Stat 1991; 19: 1–67.
37. Chen T and Guestrin C. Xgboost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.
38. Breiman L Random forests. Mach Learn 2001; 45: 5–32.
39. Jones IM, Moore DH, Thomas CB, et al. Factors affecting HPRT mutant frequency in T-lymphocytes of smokers and nonsmokers. Cancer Epidemiol Prevent Biomark 1993; 2: 249–260.
40. Bell DA, Liu Y and Cortopassi GA. Occurrence of bcl-2 oncogene translocation with increased frequency in the peripheral blood ofheavy smokers. JNCI: J Nat Cancer Inst 1995; 87: 223–224. [PubMed: 7707410]
41. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias inillumina infinium 450K DNA methylation data. Bioinformatics 2013; 29: 189–196. [PubMed: 23175756]
42. Morris TJ, Butcher LM, Feber A, et al. ChAMP: 450K chip analysis methylation pipeline. Bioinformatics 2014; 30: 428–430. [PubMed: 24336642]
43. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformat 2012; 13: 1–16.
44. Houseman EA, Molitor J and Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics 2014; 30: 1431–1439. [PubMed: 24451622]
45. Holm S A simple sequentially rejective multiple test procedure. Scandinavian J Stat 1979; 65–70.

46. Grieshober L, Graw S, Barnett MJ, et al. AHRR methylation in heavy smokers: associations with smoking, lung cancer risk, and lung cancer mortality. BMC Cancer 2020; 20.
47. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun 2015; 6: 10192. [PubMed: 26667048]
48. Zhang Y, Elgizouli M, Schöttker B, et al. Smoking-associated DNA methylation markers predict lung cancer incidence. Clin Epigenet 2016; 8: 127.
49. Bojesen SE, Timpson N, Relton C, et al. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. Thorax 2017; 72: 646–653. [PubMed: 28100713]
50. Battram T, Richmond RC, Baglietto L, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. Int J Epidemiol 2019; 48: 1493–1504. [PubMed: 31549173]
51. Díaz I and van der Laan MJ. Population intervention causal effects based on stochastic interventions. Biometrics 2012; 68: 541–549. [PubMed: 21977966]
52. Hejazi NS, van der Laan MJ, Janes HE et al. Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. Biometrics 2020; 77: 1241–1253. [PubMed: 32949147]
53. Díaz I and Hejazi NS. Causal mediation analysis for stochastic interventions. J R Stat Soc: Ser B (Statistical Methodology) 2020; 82: 661–683.
54. Hejazi NS, Rudolph KE, van der Laan MJ, et al. Nonparametric causal mediation analysis for stochastic interventional (in)direct effects. Biostatistics 2022. in press.

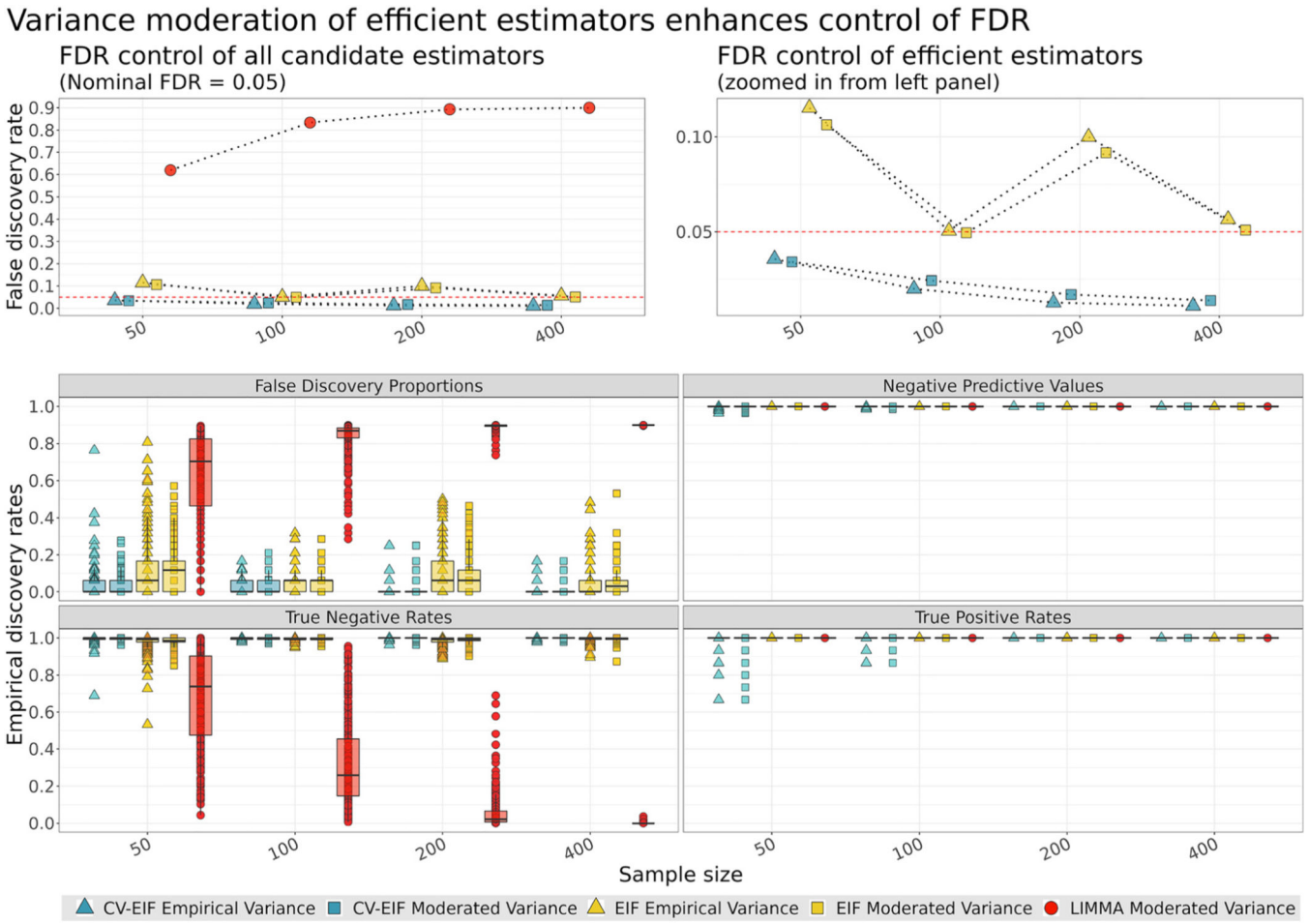

Figure 1.

Control of the False Discovery Rate (FDR) across hypothesis testing procedures in a setting with strong exposure effect in 10% of biomarkers and no positivity issues in the exposure mechanism. *Upper panel*: Control of the FDR using the Benjamini-Hochberg correction. *Lower panel*: Empirical distributions of false discovery proportions and negative predictive values, as well as of the true positive and true negative rates.

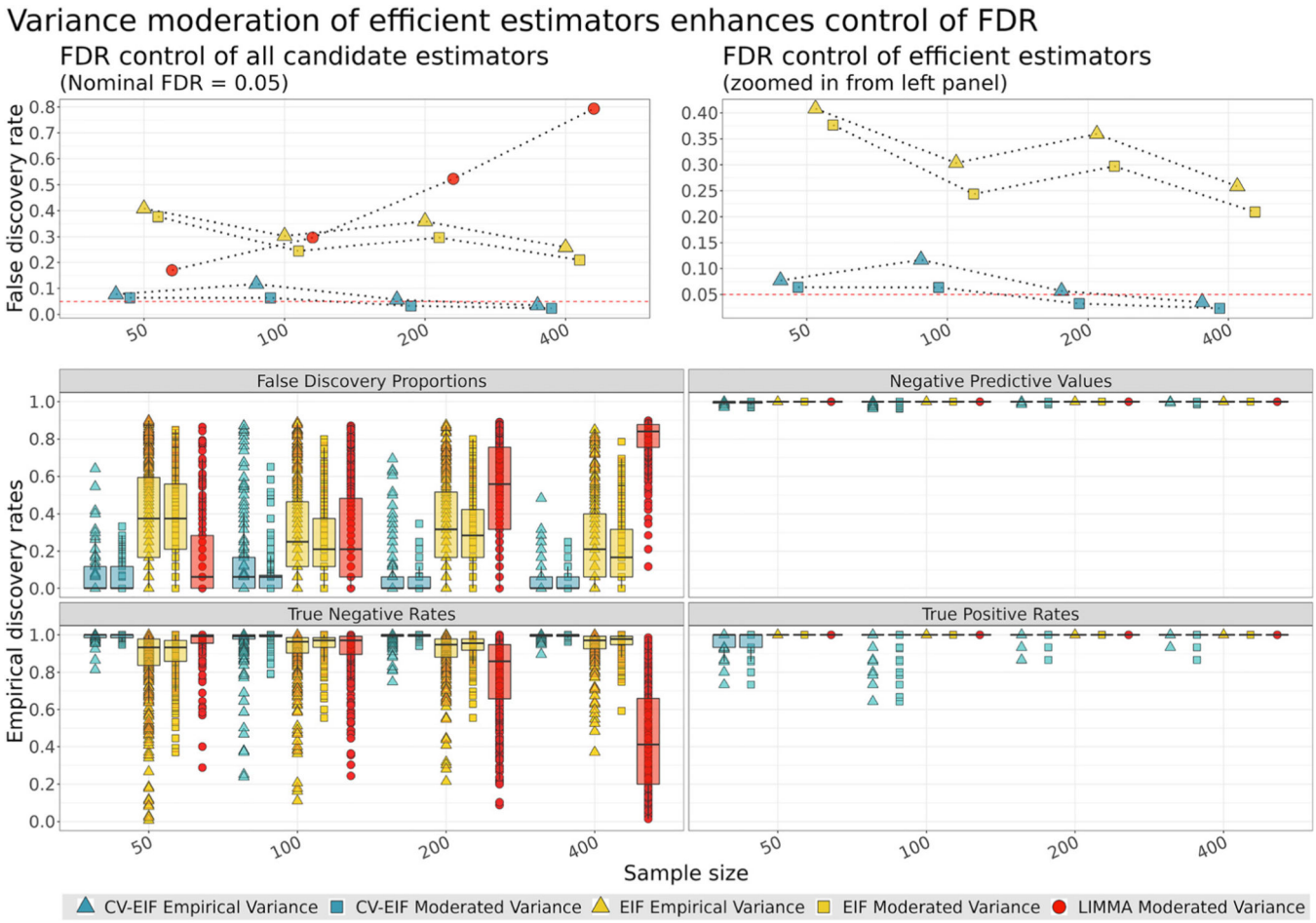

Figure 2.

Control of the False Discovery Rate (FDR) across hypothesis testing procedures in a setting with strong exposure effect in 10% of biomarkers and notable positivity issues in the exposure mechanism. *Upper panel*: Control of the FDR using the Benjamini-Hochberg correction. *Lower panel*: Empirical distributions of false discovery proportions and negative predictive values, as well as of the true positive and true negative rates.

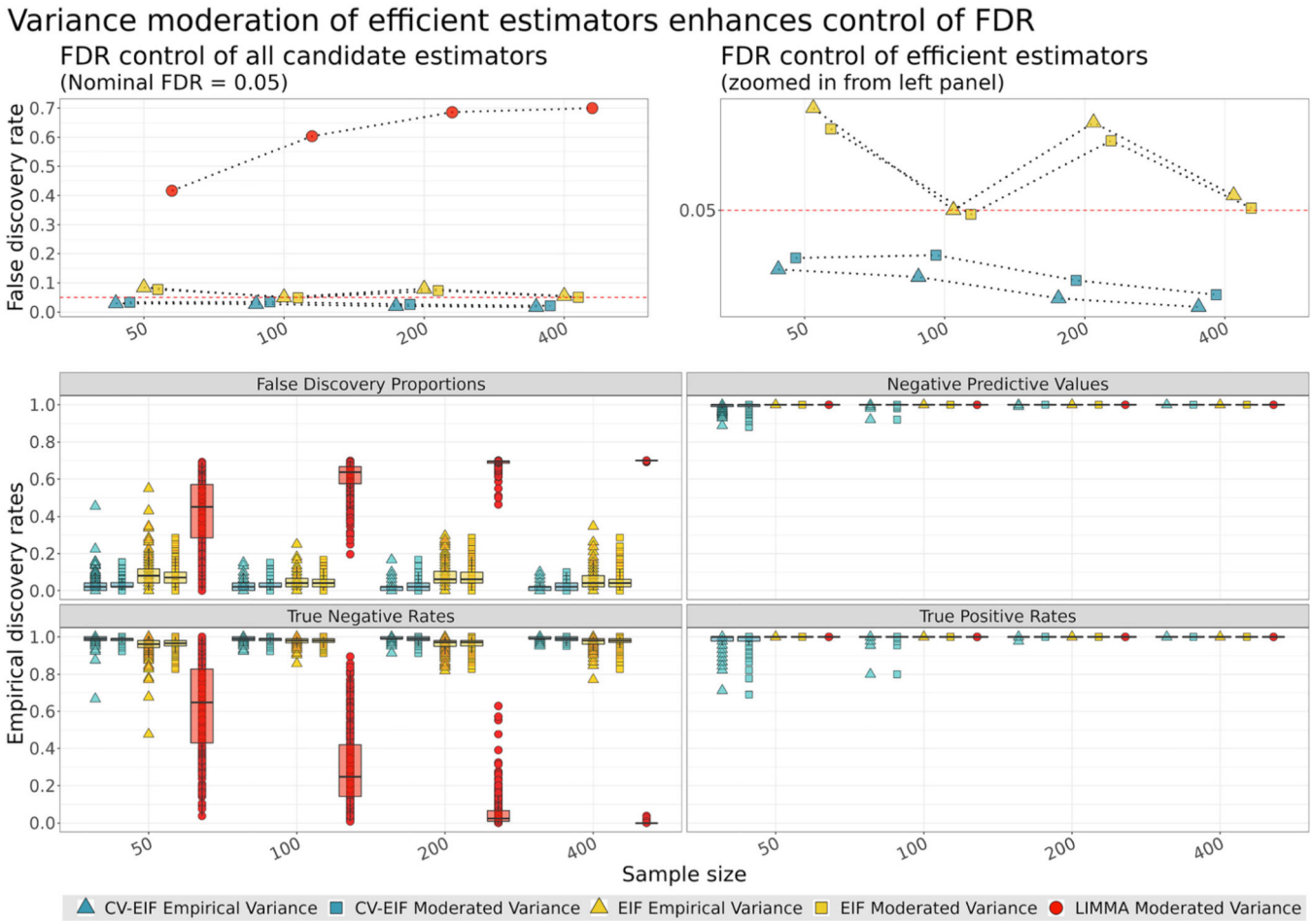

Figure 3.

Control of the False Discovery Rate (FDR) across hypothesis testing procedures in a setting with strong exposure effect in 30% of biomarkers and no positivity issues in the exposure mechanism. *Upper panel*: Control of the FDR using the Benjamini-Hochberg correction. *Lower panel*: Empirical distributions of false discovery proportions and negative predictive values, as well as of the true positive and true negative rates.

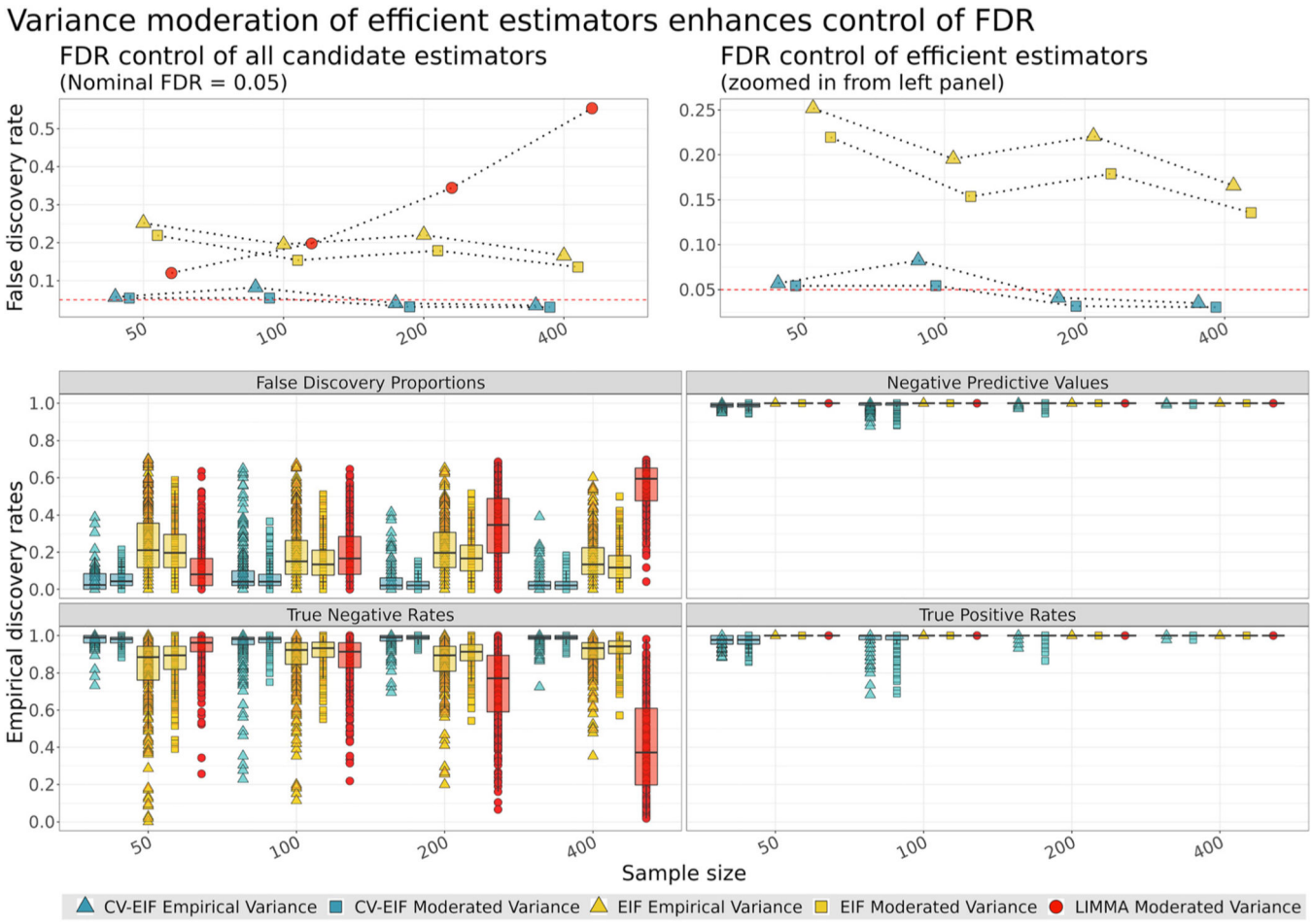

Figure 4.

Control of the False Discovery Rate (FDR) across hypothesis testing procedures in a setting with strong exposure effect in 30% of biomarkers and notable positivity issues in the exposure mechanism. *Upper panel*: Control of the FDR using the Benjamini-Hochberg correction. *Lower panel*: Empirical distributions of false discovery proportions and negative predictive values, as well as of the true positive and true negative rates.

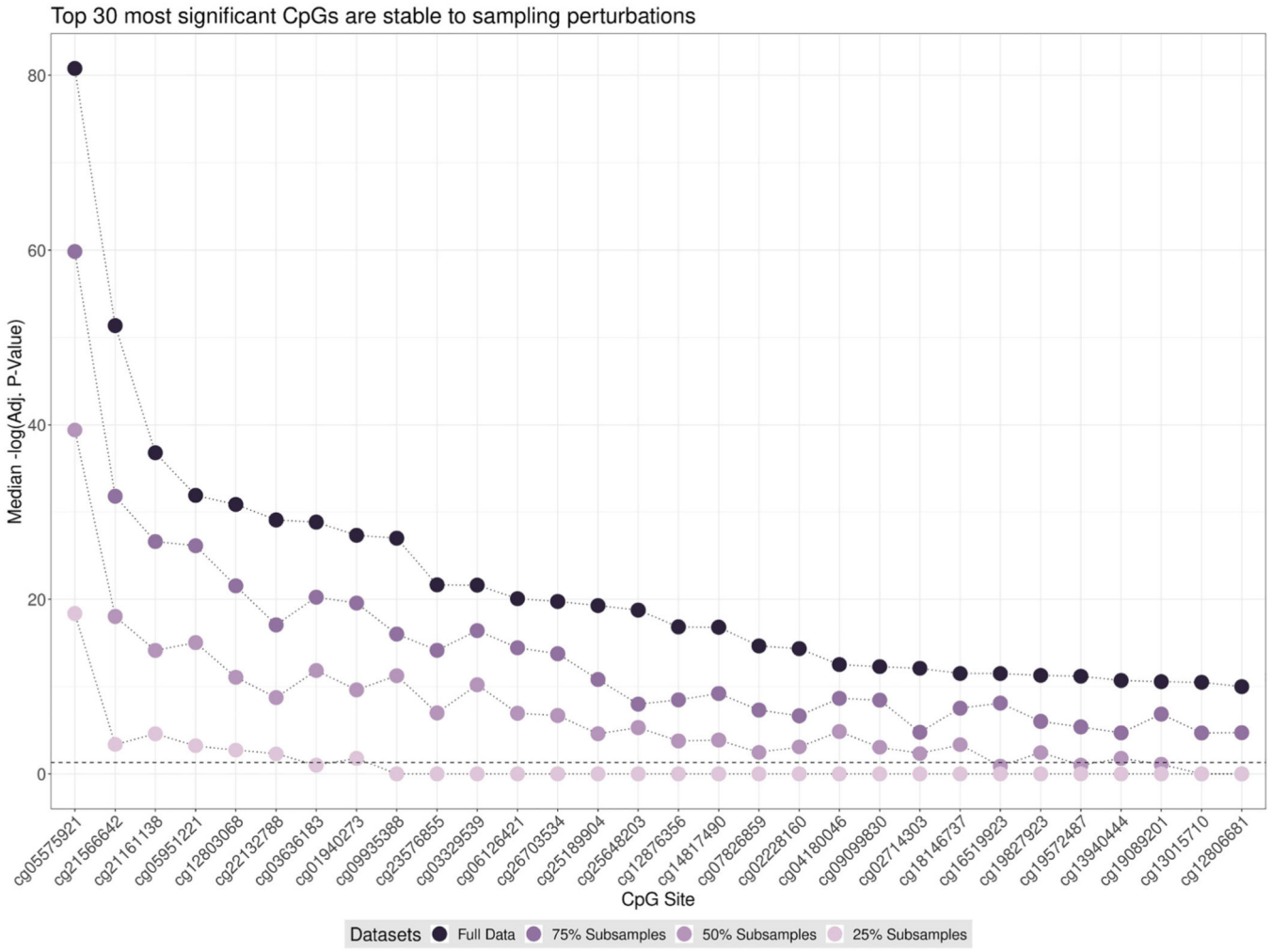

Figure 5.

Evaluation of the top 30 differentially methylated CpGs (orderd left to right) from the complete analysis in terms of median $\{-\log_{10}(\text{adj. } p-\text{value})\}$'s across the three subsampling schemes.