Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Causal mediation analysis with one or multiple mediators: a comparative study

Judith Abécassis*[1] | Julie Josse[2] | Bertrand Thirion[1]

[1]Inria Paris-Saclay, CEA, Université Paris-Saclay, Palaiseau, France
[2]Inria Sophia-Antipolis, Montpellier, France

Correspondence
*Judith Abécassis. Email: judith.abecassis@inria.fr

Summary

Mediation analysis breaks down the causal effect of a treatment on an outcome into an indirect effect, acting through a third group of variables called mediators and a direct effect, operating through other mechanisms. We provide a thorough evaluation of estimators for direct and indirect effects in the context of causal mediation analysis for binary, continuous and multi-dimensional mediators. We consider standard parametric implementations of classical estimators, and propose and assess the relevance of several extensions inspired from double or debiased machine learning, in particular non-parametric models, regularization, probability calibration and cross-fitting. Our results show that most methods obtain reasonable estimates under model misspecification, but some methods, including multiply-robust methods, are very sensitive to (near-)violations of the overlap assumption. This trend is even more pronounced in multi-dimensional settings. We also describe settings where the use of more complex non-parametric models for estimation is relevant.

To illustrate the considered methods on real data, we examine the causal path from higher education graduation to middle-age general intelligence in the UK Biobank, which includes several potential binary, continuous and multi-dimensional mediators. This analysis shows that this effect is partially mediated by having a physical occupation, and brain characteristics measured through MRI, but not by the brain age, a popular MRI-derived phenotype.

KEYWORDS:
Causal inference, mediation analysis, machine learning, cross-fitting, multi-dimensional mediators

1 | INTRODUCTION

Causal inference in observational studies is primarily used to measure the causal effect of a treatment on an outcome. Nevertheless, in a lot of fields, disentangling the mechanism of action is just as important, as it allows us to identify potential intermediate intervention targets, and more generally, deepen our understanding of the processes that lead to the observed outcome.

Causal mediation analysis aims at separating the (total) causal effect into two components: an indirect effect through a third (group of) variable(s) called mediator(s), and a direct effect through alternative path(s)[1]. A central issue solving this question is confounding, as even in the "ideal" case of a randomized controlled trial, only the exposure is randomized. Further control on the mediator is thus necessary to identify the direct and indirect effects, as outlined by[2] in the potential outcome framework[3]. Earlier work on mediation was mostly based on parametric structural equations[4,5,6], and neglected identifiability assumptions. Further

work has clarified identification assumptions, and developed new estimation approaches in parametric and non-parametric settings[2,1,7,8,9,10,11,12,13,14,15,16]. This has led to the development of more complex estimators, that proceed in two steps: first the fitting of classification or regression models on observed data, and then the use of those models' predictions to compute the effects of interest. The models of the first step are called nuisance models, as they are not directly of use, and are generally parametric linear models. Yet, the consistency of the different estimators relies on the consistency of at least a subset of those nuisance models. When the parametric model fails to represent the actual underlying phenomenon, this is called model misspecification, and can lead to erroneous prediction.

Another limitation of classical estimators is that most methods are dedicated to the case of a one-dimension binary mediator, while the problem of handling several mediators is increasingly considered in the literature.[17,18,19] have specified identifiability assumptions for direct and indirect effects, either jointly through all mediators, or for a path-specific effect through one particular mediator. An additional difficulty lies in potential causal relations between mediators[17,18,20,21]. Recent work[22,23,24] has focused on high dimension settings, where there exist a lot of mediators, potentially highly correlated, such as gene expression or medical imaging. In [24], dimensionality reduction is performed first, using Principal Component analysis (PCA) while in[23], Independent Component Analysis (ICA) or a sparse version of PCA is sued instead; the objective being to compute a smaller number of orthogonal "directions of mediation". Alternatively, [25,26] propose statistical tests to select only a relevant subset of mediators. One then conducts an analysis of mediation using structural equations models with the classic approach of the product of coefficients[6,27], which relies on a linear parametric relation between the directions of mediation and the treatment, and between the directions of mediation and the outcome.

In this work, we focus on the problem of estimating the direct and indirect effects of one or several mediators jointly (no path-specific effect) using classical approaches, or more recent approaches that are robust to model misspecification. We propose new variants of existing estimators, with more flexible machine learning models to account for complex relations between the variables of interest. We provide a comprehensive evaluation of classical and more recent methods on simulated data, extending the work of[28] for a binary mediator to the continuous and multi-dimensional mediators. We rely on a diversity of simulation settings to explore the practical implications of violations of parametric model specifications, violations to the overlap assumption, variations in the number of observations, and choice of the confounder and mediator variables. This benchmark provides a good overview of available estimators, their validity conditions and limitations which constitutes a valuable guide to the practitioner.

To go beyond performance analysis on simulated data, we conduct several mediation analyses on real data from the UK Biobank to explore cognitive functions in a cohort of middle-aged adults. UK Biobank is a prospective cohort of about 500,000 healthy participants in the UK with very thorough socio-demographic, medical, lifestyle, physical and cognitive assessment. A subset of nearly 40,000 participants also underwent a more enhanced functional exploration including brain structural and functional magnetic resonance Imaging (MRI). This unprecedentedly large imaging database allows us to assess potential role of the brain structure in the shaping of cognitive functions, while observing potential confounders. The results obtained for several potential mediators of different nature (binary, continuous and multidimensional) further illustrate the properties of the different considered estimators.

The rest of the article is organized as follows. Section 2 formalizes the causal mediation analysis problem, with the definition of the natural direct and indirect effects, and the associated identifiability assumptions. In Section 3, we present the estimators evaluated, as well as the used underlying models and implementations. Section 4 presents in details the simulation process and the main trends. An application of mediation analysis to decipher some aspects of the effect of education on middle-age cognitive functions is proposed in Section 5. Finally, we discuss the overall results in Section 6.

2 | PROBLEM SETTING: CAUSAL MEDIATION ANALYSIS

In this section, we introduce the potential outcome framework to define the causal quantities of interest. We then specify the required assumptions to identify those quantities, i.e. estimate them with the data at our disposal.

2.1 | Natural direct and indirect effect

The objective of mediation analysis is to quantify the part of the total effect achieved through the mediator, i.e. the indirect effect, and the effect of the treatment without further intermediate, i.e. the direct effect. For each individual, we denote the (binary) treatment T, the observed outcome Y, the mediator(s) M, and the covariate(s) X; covariates associated with at least two

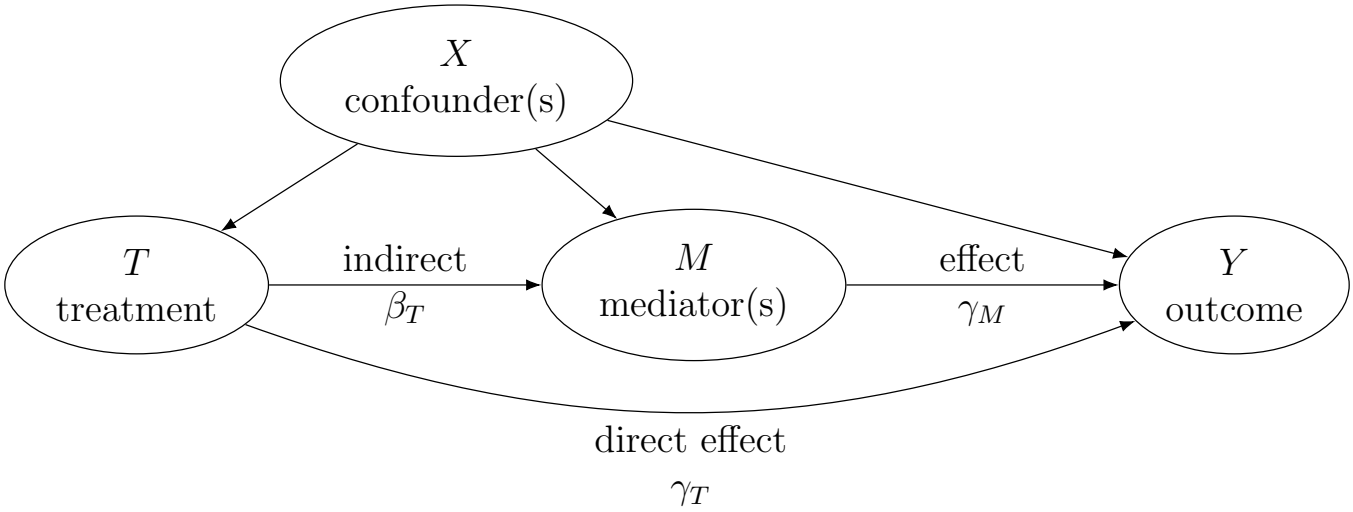

FIGURE 1 A general directed acyclic graph for mediation in causal inference. Each node represents a group of variables, and arrows denote causal relations between them. The arrows also indicate the values of the natural direct and indirect effects as specified in the structural equation model described in paragraph 3.1.1 and in the simulations.

variables among the treatment, mediator(s) and outcome are confounders and should be adjusted for. Let us note $\mathcal{T}, \mathcal{Y}, \mathcal{M}, \mathcal{X}$ the respective supports of T, Y, M and X. The relations between those variables are illustrated in Figure 1. Extending the potential outcomes framework[3], we can define $M(t)$ and $Y(t, M(t))$ the potential mediator state and the potential outcome under the treatment value $t \in \{0, 1\}$. For each unit, only one potential outcome and mediator state are observed. To define the unobserved (counterfactual) outcomes and mediators, we assume that the observed outcome and mediator are the potential outcome and mediator under the actual assigned treatment.

Assumption 1 (SUTVA (Stable Unit Treatment Values) and consistency)**.**

$$M = TM(1) + (1 - T)M(0) \text{ and } Y = TY(1, M(1)) + (1 - T)Y(0, M(0)) \tag{1}$$

We then define the total average treatment effect (ATE) as:

Definition 1 (Total average treatment effect)**.**

$$\tau = \mathbb{E}[Y(1, M(1)) - Y(0, M(0))] \tag{2}$$

To further decompose the total effect into natural direct and indirect effects[1,20,21], we define cross-world potential outcomes that correspond to varying the treatment, while maintaining the value of the mediator to the value it would have without changing the treatment, and the opposite. The word "natural" is used to distinguish those effects from the controlled direct effect, also defined by[1], which consists in measuring the direct effect while intervening to artificially set the mediator to a fixed value. We will not consider the estimation of the controlled direct effect in this study. Contrary to the previously mentioned potential outcomes, where one of them is observed, cross-world outcomes can never be observed. Those additional terms allow us to define the natural direct effect as

Definition 2 (Natural direct effect)**.**

$$\theta(t) = \mathbb{E}[Y(1, M(t)) - Y(0, M(t))], \quad t \in \{0, 1\},$$

and the natural indirect effect

Definition 3 (Natural indirect effect)**.**

$$\delta(t) = \mathbb{E}[Y(t, M(1)) - Y(t, M(0))], \quad t \in \{0, 1\}.$$

The ATE in equation (2) is the sum of the direct and indirect effects of opposite treatment states: $\tau = \theta(0)+\delta(1) = \theta(1)+\delta(0)$, resulting in well-defined effect decomposition.

2.2 | Identification assumptions

As mentioned above, the parameters of interest include both unobserved and unobservable terms. In addition to the SUTVA and consistency assumptions, already introduced in the previous paragraph, identification requires additional assumptions:

Assumption 2 (Sequential conditional independence of the treatment[9])**.**

$$\{Y(t', m), M(t)\} \perp\!\!\!\perp T|X \quad \text{for all } t', t \in \{0, 1\} \text{ and } m \in \mathcal{M}, \tag{3}$$

where $\perp\!\!\!\perp$ denotes statistical independence.

Assumption 2 imposes the absence of unobserved confounders of the treatment and the outcome on the one hand, and of the treatment and the mediator on the other hand.

Assumption 3 (Sequential conditional independence of the mediator[9])**.**

$$Y(t', m) \perp\!\!\!\perp M|T = t, X = x \quad \text{for all } t', t \in \{0, 1\}, m \in \mathcal{M} \text{ and } x \in \mathcal{X},$$

Assumption 3 forbids the existence of unobserved confounders affecting both the mediator and the outcome.

Finally, a common support assumption (also called positivity or overlap) is needed

Assumption 4 (Positivity assumption)**.**

$$\mathrm{P}(T = t|X = x) > 0 \text{ and } 0 < \mathrm{P}(M(t) = m|T = t, X = x) \text{ for all } t', t \in \{0, 1\},\ m \in \mathcal{M} \text{ and } x \in \mathcal{X}.$$

Relying on Assumptions 2, 3, and 4 total effect, along with natural direct and indirect effects are identifiable. We give the demonstration in the Appendix section S1 that the mean potential outcomes and cross-world potential outcomes needed to compute the effects of interest can be identified non-parametrically, that is without any form restriction such as linearity.

2.3 | The case of several mediators

If we now consider several mediators of interest, $\mathbf{M} = (M^{(1)}, \ldots, M^{(K)})$, and aim at computing the indirect effect through all the mediators jointly, all the definitions and assumptions above can be written similarly, by just replacing the mediator M by a mediator vector $\mathbf{M}$. However, identification of path-specific effects through one particular mediator requires additional assumptions[17,20] and is beyond the scope of this work.

3 | OVERVIEW OF ESTIMATORS FOR MEDIATION ANALYSIS

In this study, we consider several estimators for the natural direct and indirect effects that apply to binary, continuous and/or multi-dimensional mediators. We compare their properties, conditions of application and performances on simulated data with binary and continuous multi-dimensional mediators in Section 4. This section introduces the different estimators, and provides some details on the implementations used in the experiments. Some estimators embed a procedure for uncertainty estimation but not all, so we restricted our comparison solely to the estimation of direct and indirect effects, and assess the variance of estimators with several independent draws of data. We consider a sample of n $i.i.d.$ observations denoted $(X_i, T_i, M_i, Y_i), i \in \{1, \ldots, n\}$.

3.1 | Parametric estimators

3.1.1 | Coefficient product

The method of coefficient product developed by[6] is the first developed estimator for mediation analysis. It consists in assuming the following structural equation model:

$$M(x, t) = \beta_0 + \beta_T t + x^T \beta_X + \epsilon_M \tag{4}$$

$$Y(x, t, m) = \gamma_0 + \gamma_T t + x^T \gamma_X + \gamma_M m + \epsilon_Y \tag{5}$$

where ϵ_M and ϵ_Y are independent centered normal random variables.

Then we have, after estimation of the regression parameters

$$\begin{aligned}\hat{\theta}(0) &= \hat{\theta}(1) = \hat{\gamma_T}\\ \hat{\delta}(0) &= \hat{\delta}(1) = \hat{\gamma_M}\hat{\beta_T}\\ \hat{\tau} &= \hat{\theta}(1) + \hat{\delta}(0) = \hat{\theta}(0) + \hat{\delta}(1).\end{aligned}$$

This estimator is consistent if the model for the outcome and for the mediator are both correctly specified. The model can be easily extended to binary mediators or outcomes using logistic regressions, or to situations with interactions by adding interaction terms to the structural equations[29].

3.1.2 | g-Computation

The following formula is demonstrated in section S1, where we denote $f_{A=a}$ and $f_{A=a|B=b}$ the probability density function of a random variable A, unconditionally or conditionally given other random variable(s) B:

$$\mathbb{E}\left[Y\left(t, M\left(t'\right)\right)\right] = \iint \mathbb{E}[Y|T=t, M=m, X=x] f_{M=m|T=t',X=x} f_{X=x} dm dx$$

It directly yields the mediation formula[1,9],

$$\begin{aligned}\theta(t) &= \iint \{\mathbb{E}[Y \mid T=1, M=m, X=x] - \mathbb{E}[Y \mid T=0, M=m, X=x]\} f_{M=m|T=t,X=x} dm f_{X=x} dx\\ \delta(t) &= \iint \mathbb{E}[Y \mid T=t, M=m, X=x] \left\{f_{M=m|T=1,X=x} - f_{M=m|T=0,X=x}\right\} dm f_{X=x} dx.\end{aligned}$$

In practice, we perform parametric or non-parametric estimation of $\hat{\mu}_Y(t,m,x)$ for the conditional mean outcome $\mathbb{E}[Y \mid T = t, M = m, X = x]$ and $\hat{f}(m|t,x)$ for the conditional mediator density $f_{M=m|T=t,X=x}$ (or conditional probability if the mediator is discrete), and the final effects are estimated as follows

$$\begin{aligned}\hat{\theta}(t) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=0}^{1}\left\{(\hat{\mu}_Y(1,m,X_i) - \hat{\mu}_Y(0,m,X_i))\hat{f}(m|t,X_i)\right\},\\ \hat{\delta}(t) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{m=0}^{1}\left\{\hat{\mu}_Y(t,m,X_i)(\hat{f}(m|1,X_i) - \hat{f}(m|0,X_i))\right\}.\end{aligned}$$

The g-Computation estimator is consistent if both plug-in models are correctly specified. The estimation of the conditional mediator density is challenging for multi-dimensional mediators. Indeed, one can perform either parametric density estimation with strong assumptions on the distribution and the variance-covariance matrix, or non-parametric density estimation, which requires a number of observations that increases exponentially with the dimension[30].

3.1.3 | Simulation-based estimator

Simulation-based estimation[31] consists in simulating unobserved potential outcomes and cross-world potential outcomes using fitted estimators for the outcome and the mediator model to directly compute the indirect and direct effects. In current implementation, such a method is implemented with linear models.

3.2 | Semi-parametric estimators

3.2.1 | Inverse probability weighting estimator (IPW)

The identifiability computation in step (A3) in Appendix S1 provides the ground for an inverse probability weighting-type approach (IPW), developed in[13].

Computation requires the estimation of two nuisance parameters, the conditional probability of treatment given covariates X $p(X) = \mathbb{P}(T = 1 \mid X)$, and the conditional probability of treatment given covariates and mediator(s) $\rho(X) = \mathbb{P}(T = 1 \mid X, M)$.

In practical implementation, the weights are normalized for more stability; we can write the normalized estimator as follows

$$\hat{\theta}(0) = \frac{\sum_{i=1}^{n} Y_i T_i (1-\hat{\rho}(M_i, X_i))/[\hat{\rho}(M_i, X_i)(1-\hat{p}(X_i))]}{\sum_{i=1}^{n} T_i (1-\hat{\rho}(M_i, X_i))/[\hat{\rho}(M_i, X_i)(1-\hat{p}(X_i))]} - \frac{\sum_{i=1}^{n} Y_i (1-T_i)/(1-\hat{p}(X_i))}{\sum_{i=1}^{n} (1-T_i)/(1-\hat{p}(X_i))}$$

$$\hat{\delta}(1) = \frac{\sum_{i=1}^{n} Y_i T_i / \hat{p}(X_i)}{\sum_{i=1}^{n} T_i / \hat{p}(X_i)} - \frac{\sum_{i=1}^{n} Y_i T_i (1-\hat{\rho}(M_i, X_i))/[\hat{\rho}(M_i, X_i)(1-\hat{p}(X_i))]}{\sum_{i=1}^{n} T_i (1-\hat{\rho}(M_i, X_i))/[\hat{\rho}(M_i, X_i)(1-\hat{p}(X_i))]}$$

The IPW estimator is consistent if parametric models for p and ρ are consistently estimated. The absence of estimation of the density of the estimator(s) allows us to use this estimator when the mediator is continuous and/or multi-dimensional. However, the total effect does not involve $\mathbb{P}(T = 1 \mid X, M)$, hence it is robust to misspecification of the mediator model.

3.2.2 | Multiply-robust estimator

The inverse mediator density weighting implemented in the multiply-robust estimator[12] is motivated by the last step A4 of the identifiability computation in Appendix S1.

The estimator is derived using the sample analogue of the efficient influence function for computing potential outcomes. We can then write[12,28]

$$\begin{aligned}\hat{\theta}(0) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{ &\left[\frac{T_i \hat{f}(M_i|0, X_i)}{\hat{p}(X_i)\hat{f}(M_i|1, X_i)} - \frac{1-T_i}{1-\hat{p}(X_i)}\right] \times (Y_i - \hat{\mu}_Y(T_i, M_i, X_i)) \\ &+ \frac{1-T_i}{1-\hat{p}(X_i)} \times (\hat{\mu}_Y(1, M_i, X_i) - \hat{\mu}_Y(0, M_i, X_i) - \hat{\theta}(0, X_i)) + \hat{\theta}(0, X_i)\Bigg\} \end{aligned} \tag{6}$$

$$\begin{aligned}\hat{\delta}(1) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{ &\frac{T_i}{\hat{p}(X_i)}\left[Y_i - \hat{\psi}(1, X_i) - \frac{\hat{f}(M_i|0, X_i)}{\hat{f}(M_i|1, X_i)}(Y - \hat{\mu}_Y(1, M_i, X_i))\right] \\ &- \frac{1-T_i}{1-\hat{p}(X_i)}(\hat{\mu}_Y(1, M_i, X_i) - \hat{\psi}(0, X_i)) + \hat{\psi}(1, X_i) - \hat{\psi}(0, X_i)\Bigg\} \end{aligned} \tag{7}$$

with $\hat{\mu}_Y(t, m, x)$ estimating the conditional mean outcome $\mathbb{E}[Y|T = t, M = m, X = x]$, $\hat{f}(m|t, x)$ estimating the conditional mediator density $f_{M|T=t,X=x}(m)$ (or conditional probability if the mediator is discrete), $\hat{p}(x)$ estimating the treatment propensity score $P(T = 1|X = x)$, $\hat{\theta}(t, x)$ estimating the conditional direct effect given X $\mathbb{E}_{M|T=t,X=x}[\mathbb{E}[Y|T = 1, M = m, X = x] - \mathbb{E}[Y|T = 0, M = m, X = x]|T = t, X = x]$ which is obtained by regressing $\hat{\mu}_Y(1, M, X) - \hat{\mu}_Y(0, M, X)$ on X among those with treatment $T = t$ and $\hat{\psi}(t, x)$ estimating $\mathbb{E}_{M|T=t,X=x}[\mathbb{E}[Y|T = 1, M = m, X = x]|T = t, X = x]$.

A very interesting property of this estimator is that it is triply-robust, as it remains consistent if at least two of the three following models are well specified: (i) the conditional mean outcome, $\mathbb{E}[Y|T, M, X]$, (ii) the conditional density of M given T, X, and (iii) the treatment propensity score. This estimator is also efficient (under the nonparametric model) if (i), (ii), and (iii) are all correctly specified. The estimation of the mediator conditional density prevents the use of this estimator for a continuous and/or multi-dimensional mediator.

3.2.3 | G-estimator

In[16], the authors propose an estimator in the case where models for the conditional expectation of the mediator and outcome are partially linear. The authors consider

$$\mathbb{E}[M|T, X] = \beta_1 T + f(X) \tag{8}$$

$$\mathbb{E}[Y|T, M, X] = \beta_2 M + \beta_3 T + g(X) \tag{9}$$

where f and g are arbitrary functions. The target parameters are $\beta = (\beta_1, \beta_2, \beta_3)$, while f, g, and $p(X) = \mathbb{E}[T|X]$ are nuisance parameters. Both target and nuisance parameters are estimated jointly using an alternating estimation procedure. The estimation of target parameters relies on the product of residuals after regressing out the terms with f, g, p. For nuisance parameters estimation, a bias-reduction strategy is implemented[32,33].

The G-estimator is consistent if models 8 and 9 hold, and if two out of three models for f, g, p are consistently estimated. There is no estimation of the mediator conditional density, so the G-estimator can easily handle a continuous mediator, but not a multi-dimensional mediator.

3.2.4 | Double-Machine Learning for mediation estimator (medDML)

The multiply-robust estimator was extended in [15], by removing the requirement to estimate the conditional mediator density using Bayes' law. Indeed, the authors demonstrate that

$$\frac{f(M|T=1-t,X)}{p(T=t|X)f(M|T=t,X)} = \frac{(1-p(T=t|X,M))f(M|X)}{1-p(T=t|X)} \frac{p(T=t|X)}{p(T=t|X,M)f(M|X)p(T=t|X)}$$
$$= \frac{1-p(T=t|X,M)}{p(T=t|X,M)(1-p(T=t|X))}$$

This new expression can be directly used in equations (6) and (7), lifting the previously mentioned restrictions for continuous and/or multi-dimensional mediators. Additionally, the authors show that this new estimator is Neyman orthogonal, which is crucial to the application of double machine learning, in addition to the use of cross-fitting [34].

3.3 | Implementation considerations

We re-implemented some of the considered estimators, using Python package `scikit-learn` [35]. This re-implementation has allowed us more flexibility in the choice of algorithms for nuisance parameters estimation, in particular the use of random forests [36]. We have also implemented variations around those estimators using regularization, probability calibration [37,38] and cross-fitting [34]. For the other estimators, we used existing implementation in R using the `rpy2` package. The main characteristics of available implementations for each estimator are presented in Table 1.

estimator	binary	continuous	multi dimensional	reference R implementation	python re-implementation
coefficient product 3.1.1	x	x	x	no	linear regression for both the mediator and the outcome models with a very small L2 regularization ($\alpha = 10^{-5}$)
g-Computation 3.1.2	x	-	-	no	linear and logistic regression (with or without L2 regularization) or random forests to estimate $\hat{\mu}_Y$ and $\hat{f}$
simulation-based estimator 3.1.3	x	x	-	function `mediate` in the R package `mediation` [39], with parametric linear models for the mediator and outcome models	no
IPW 3.2.1	x	x	x	function `medweight` in the R package `causalweight` [40] with logit or probit regression to estimate conditional probabilities	logistic regression (with or without L2 regularization) or random forests to estimate conditional probabilities
multiply- 3.2.2 robust estimator	x	-	-	no	linear models (with or without L2 regularization) or non-parametric estimators (random forests)
g-estimator 3.2.3	x	x	-	function `G_estimation` from the `plmed` R package (http://github.com/ohines/plmed) [16]	no
medDML 3.2.4	x	x	x	function `medDML` in the R package `causalweight` [40], with no trimming, and default parameters for the other options	no

TABLE 1 Characteristics of estimators and their implementations The columns "binary", "continuous" and "multi-dimensional" refer to the ability of methods to handle mediator of this type. Symbols "x" and "-" mean possible or impossible respectively. Most estimators have available implementations in R. We have reimplemented some of them in Python to allow more flexibility in the choice of nuisance models. Our implementations are available as a Python package at https://github.com/judithabk6/med_bench.

4 | PERFORMANCES ON SIMULATIONS

We assess the performances of the estimators presented in the previous section using simulated data. We tried to cover a variety of situations to establish the strengths and limitations of each approach. We first present our simulation settings, and then our main findings.

4.1 | Generation of simulated data

We have generated simulated datasets with a number of varying characteristics: the number of observations $n \in \{500, 1000, 10000\}$, the dimension of the covariates X, either 1, 5 or 20, the dimension (and variable type - binary or continuous) of the mediator M, either 1, 5 or 20, the linearity (or not) of the mediator and outcome models, the mediated proportion, through varying the treatment parameter in the mediator model.

We note K the number of dimensions of the confounder X, and K_{obs} the observed number of dimensions of X ($K_{obs} \geq K$) to reproduce the common situation where the analyst does not know the true set of confounder variables. We define a confounder as a variable causally associated with at least two the three variable types treatment, outcome and mediator. We follow guidelines from[41] and associate the (potentially disjoint) sets $T - Y$, $T - M$ and $M - Y$ confounders in a single set X. Similarly for mediators, L is the actual number of dimensions of the mediator (an actual mediator is characterized by a non-zero causal effect from the treatment on the mediator, and a non-zero effect of the mediator on the outcome), and L_{obs} the observed number of dimensions of the mediator.

We have used the following simulation framework

$$
\begin{aligned}
&X \sim \mathcal{N}(0, I_{K_{obs}}) \\
&T \sim \text{Bernoulli}(\text{expit}(\alpha^T X)) \qquad \text{with } \alpha = [\mathbf{1}_K \cdot \mathbf{0}_{K_{obs}-K}]/K \\
&\begin{cases}
M \sim \text{Bernoulli}(\text{expit}(u(\beta_X^T X) + 2w_T T)) & \text{if } M \text{ is binary and } L = L_{obs} = 1 \\
\qquad\qquad & \text{with } \beta_X = [\mathbf{1}_K \cdot \mathbf{0}_{K_{obs}-K}]/K \\
\qquad\qquad & \text{and } \gamma_M = 1 \\
M_l \sim u(\beta_X^T X) + w_T(l\%3+1)T + \mathcal{N}(0, \sigma_M^2) & \text{if } M \text{ is continuous} \\
\qquad\qquad & \text{with } \beta_X \text{ having 2 randomly picked coefficients} \\
\qquad\qquad & \text{set to 1 among the first } K_{obs} \text{ and 0 otherwise} \\
\qquad\qquad & \text{and } \gamma_M = [\mathbf{1}_L \cdot \mathbf{0}_{L_{obs}-L}] * 0.5/L
\end{cases} \\
&Y \sim u(\gamma_X^T X) + \gamma_M^T M + \gamma_T T + \mathcal{N}(0, \sigma_Y^2)
\end{aligned}
$$

where $\mathbf{1}_d$ and $\mathbf{0}_d$ are vectors of length d containing only ones and zeros respectively, $[v \cdot w]$ denotes the operation of concatenation of vectors v and w, $a\%b$ the rest of the euclidean division of a by b, I_d the identity matrix of dimension d.

The function u is introduced to add non-linearity to the model. It can be set to different values if one wants misspecification for the mediator model, the outcome model or both.

- in the linear case, u is identity,
- for misspecification, $u(x) = \frac{1}{1+e^{-x}} \quad \forall x \in \mathbb{R}$
- for severe misspecification $u(x) = 3\sin(3x) \quad \forall x \in \mathbb{R}$

The coefficient w_T is used to modulate the mediated proportion. It was set to 0.1 in the low mediated proportion setting, 1 for medium, and 5 for high.

We have considered six combinations of dimensions for covariates and mediators, three with a binary mediator in one dimension, and an increasing dimension for covariates: (i) $K = K_{obs} = L = L_{obs} = 1$, (ii) $K = K_{obs} = 5; L = L_{obs} = 1$, (iii) $K = 5; K_{obs} = 20; L = L_{obs} = 1$, and three with multidimensional continuous mediators: (iv) $K = K_{obs} = L = L_{obs} = 5$, (v) $K = 5; K_{obs} = 20; L = L_{obs} = 5$ and (vi) $K = 5; K_{obs} = 20; L = 5; L_{obs} = 20$.

For all simulations, we have set $\sigma_M = \sigma_Y = 0.5$ and $\gamma_T = 1.2$.

The true effects of interest are defined as

$$\theta(0) = \theta(1) = \gamma_T$$
$$\delta(0) = \delta(1) = \begin{cases} \mathbb{E}\left[\left(\text{expit}(f(\beta_X^T X) + \beta_T) - \text{expit}(f(\beta_X^T X)\right)\gamma_M\right] & \text{if } \boldsymbol{M} \text{ is binary} \\ \beta_T^T \gamma_M & \text{if } \boldsymbol{M} \text{ is continuous} \end{cases}$$

We generated 30 repetitions for each of the 378 generated combinations of parameters.

4.2 | Results

For readability of the figures, we present results for dimension combinations (i), (iii) and (iv), which we find representative of other results, and for linear and severe misspecifications. The "milder" misspecification exhibited results close to the linear case, indicating that important non-linearities are required to obtain inconsistent estimates for some estimators. We also have not seen significant variations in the results when including additional variables that are not confounders in the adjustment set, and similarly for adding mediators that are not actual mediators for estimation.

4.2.1 | General trends

We have applied the seven mediation estimators (with a total of 34 variant implementations) considered in this study on a variety of simulated datasets with varying dimensions of both mediators and covariates, and degree of non-linearity of the outcome and mediator models. We compare the relative error for the total, direct and indirect effect, defined as $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ. We first compare estimators using the most simple setting (non-regularized linear models in Figures 2 and 3), and then assess the contribution of more complex estimation approaches. In Figure 2, we present the errors for the total, direct and indirect effect with a "medium" mediated proportion and $n = 10,000$ observations. The total effect is almost always well estimated, but not the direct and indirect effects that exhibit opposite errors, leading to higher relative error for the indirect effect, as this effect is much smaller in magnitude. For the simplest simulations with a one-dimensional binary mediator and one-dimensional confounder, the different estimators behave as expected with very small errors in the "linear" setting where all models are well specified, small errors for the coefficient product, the g-computation and the simulation-based estimators for severe misspecification of the mediator or the outcome models. The IPW estimator exhibits a large error when the mediator model is misspecified, but is robust to outcome model misspecification. The multiply-robust estimator has a low error when either the outcome or the mediator model is misspecified, as expected. The G-estimator is surprisingly not robust to the outcome model misspecification, although the partially linear outcome model should account for the simulated deviation from the linear model. Finally, the medDML R implementation fails to produce an estimation with a one-dimensional confounder. When both the outcome and the mediator models are misspecified, all estimators fail. Results are very similar when considering confounder of dimension 20, of which only 5 dimensions are true confounders (other dimensions where not involved in the treatment, the mediator or the outcome models). We verify the multiply-robust property of the medDML estimator in this setting. Regarding the multi-dimensional setting, only three estimators work. The coefficient product method provides surprisingly good results, with no error when up to one model is misspecified, while the medDML and IPW estimators fail to estimate the direct and indirect effects in all cases.

In Figure 3, we consider the influence of the mediated proportion, defined as $\frac{\delta(t)}{\tau}$. A high mediated proportion results in a high instability of the IPW and the medDML estimators. With further exploration, the explanation lies in the prediction of propensity scores from the mediator: with a high mediated proportion, which is generated in our simulations with a high coefficient for the treatment in the mediator model, violations of the positivity assumptions can occur. The propensity score gets very close to either 0 or 1, leading to an impaired re-weighting of the samples. This problem appears even faster with multi-dimensional mediators. Interestingly, in the case where both the mediator and the outcome models are misspecified, error increases for all estimators for the direct effect estimation, but decreases for the indirect error estimation (Figure S2).

4.2.2 | Some insights on how to practically estimate nuisance parameters

We further explore the effect of some implementation variations of nuisance parameters estimation, namely the use of non-parametric models such as random forests [36] (Figure 4), the regularization of machine learning models (Figure S4), the calibration of predicted probabilities [37,38] (Figures S8 and S9), and finally the use of cross-fitting [34] (Figures S5, S6 and S7).

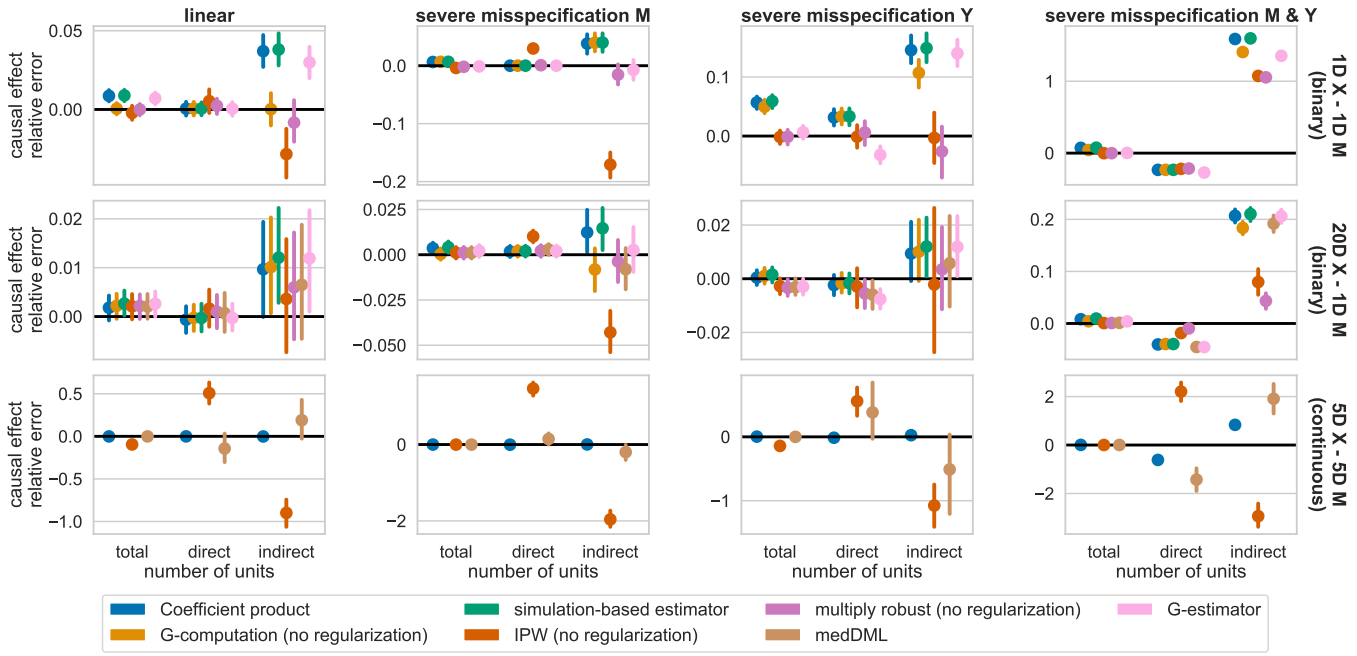

FIGURE 2 Total and natural direct and indirect effects. We show results for four scenarii of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator and covariate dimensions, labeled on the right. Each dot represents the average relative error (i.e. $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ) over 30 repetitions, and the error bars are bootstrap 95% confidence intervals. All simulations are in the "medium" mediated proportion framework with $n = 10,000$ observations. The medDML method does not accept one-dimensional covariates so we see no result for that estimator on the first row. Similarly, most estimators only handle binary one-dimensional mediators, and have no results for the third row. The total effect is generally well estimated in all situations, but not the direct and indirect effect. The indirect effect is smaller which leads to a higher relative error than for the direct effect. Model misspecifications engenders estimation errors for most estimators.

Using forests for nuisance parameters estimation (Figure 4) generally yields good results, similar to the parametric models when they are well specified, but with a slower convergence. The perform better when the models are misspecified. In the latter case, the number of samples required to obtain unbiased estimation is high. The use of regularized instead of non-regularized models for estimation with parametric models (Figure S4) has little effect. Probability calibration (Figures S8 and S9) has a slight positive effect, but the IPW estimator remains quite unstable, and calibration can either improve or damage the performance. Finally, the use of cross-fitting has little impact on error in our simulations, but requires more observations (Figures S5, S6 and S7).

5 | APPLICATION TO COGNITIVE FUNCTION IN THE UK BIOBANK

Education is generally associated with an increased intelligence[42]. However, the underlying mechanisms remain elusive. The currently admitted hypothesis is the one of the "cognitive reserve", which would rely on the structure and biological functioning of the brain through the phenomenon of neuroplasticity[43]. Yet, evidence for this hypothesis remains limited, as well as the effect size of this effect. Mediation analysis provides a very well suited framework to properly identify this effect. The UK Biobank imaging study[44] constitutes an opportunity to tackle this question, with an unprecedented cohort size with brain MRI (a proxy to brain state and current cognitive capabilities), and a very comprehensive assessment of lifestyle, physical and socio-demographic characteristics of the participants. In this work, we have retained a sub-group of 16,157 participants having undergone brain imaging.

We will consider three possible mediators: (i) a binary indicator for having a physical job, a proxy for the complexity of occupation, deemed to also have an impact on cognitive functions, (ii) brain age delta, representing the difference between the

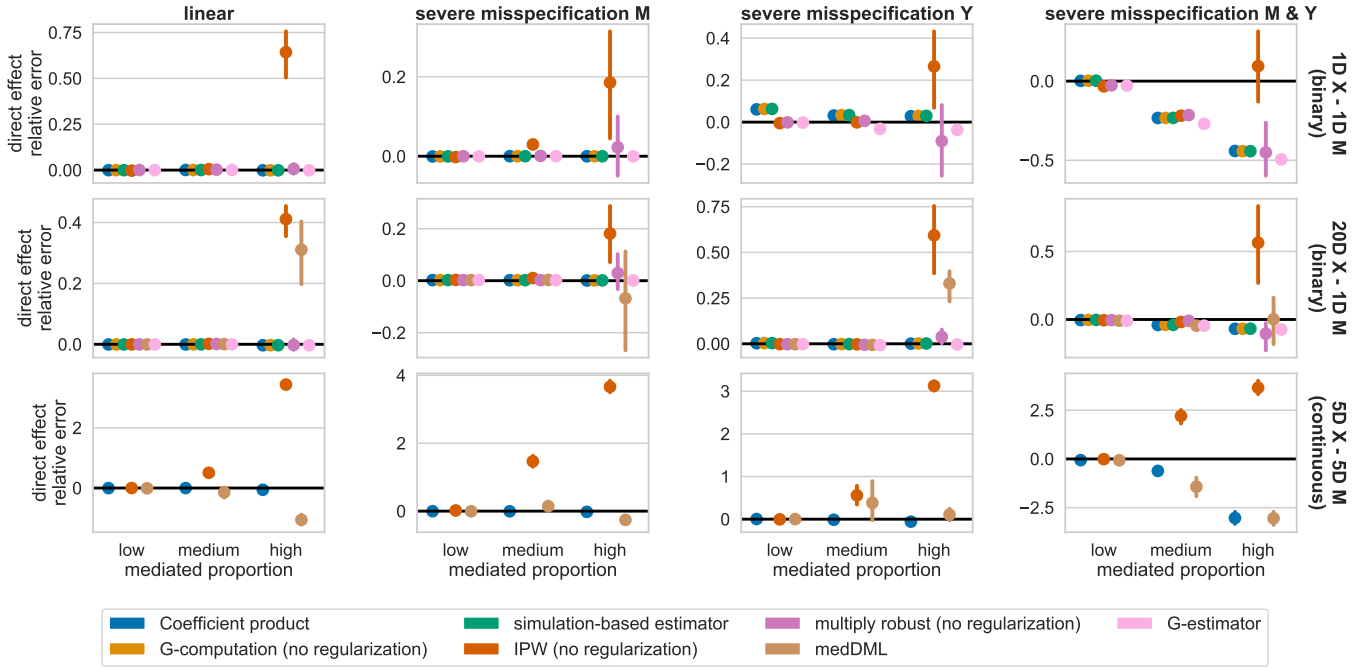

FIGURE 3 Effect of the mediated proportion on direct effect estimation. The mediated proportion is the ratio of the natural indirect effect and the total effect. As it increases, the overlap assumption can be violated. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. We show results for four scenarii of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator and covariate dimensions, labeled on the right. Each dot represents the average relative error (i.e. $\frac{\hat{\theta}-\theta}{\theta}$ for the natural direct effect θ) over 30 repetitions, and the error bars are bootstrap 95% confidence intervals. All simulations are with a number of observations $n = 10,000$. The medDML method does not accept one-dimension covariates so we see no result for that estimator on the first row. Similarly, most estimators only handle binary one-dimension mediators, and have no results for the third row. We see that the IPW and medDML estimators exhibit an important relative error in the high mediated proportion setting with the one-dimensional mediator (two first rows), and for the medium and high mediated proportions with the multi-dimensional mediators.

age predicted from brain imaging and the actual age, which is a proxy for brain health, and (iii) the first ten principal components of brain images-derived variables to represent more generally the brain characteristics. We adjust for sex, age, the center of assessment (closest to the residence), Townsend deprivation index, present and past smoking statuses, alcohol consumption frequency, early life factors (country of birth, adoption status, maternal smoking status...), and body mass index (BMI). For the brain imaging mediator (mediator (iii)), we also adjusted for head size, and head position in the MRI machine. Cognitive function was assessed by 10 distinct tests, which we summarized with dimension reduction to obtain the general intelligence factor (g-factor)[45]. The codes for all used variables are available in Supplementary Table S1.

We have applied a relevant and representative subset of the considered estimators, and present results in Figures 5, 6 and 7. The uncertainty of estimation was assessed by 15 bootstrap repetitions.

We observe no mediation through brain age delta (Figure 6), which is consistent with the recent finding that educational attainment does not influence brain aging[46], but a slight indirect effect of having a physical job (Figure 5), which is in line with the fact that occupation also influences brain reserve, and eventually a small indirect effect of brain imaging (Figure 7) which could support the potential role of neuroplasticity.

Overall, most methods provide similar results for those four problems. We observe that the G-estimator failed to provide an estimate in all cases, that medDML is highly unstable (we adjusted the axes limits to be able to see more subtle differences between the other estimators), as well as multiply robust estimator without regularization. This outlines the relevance of introducing more robust extended estimators for real-life applications.

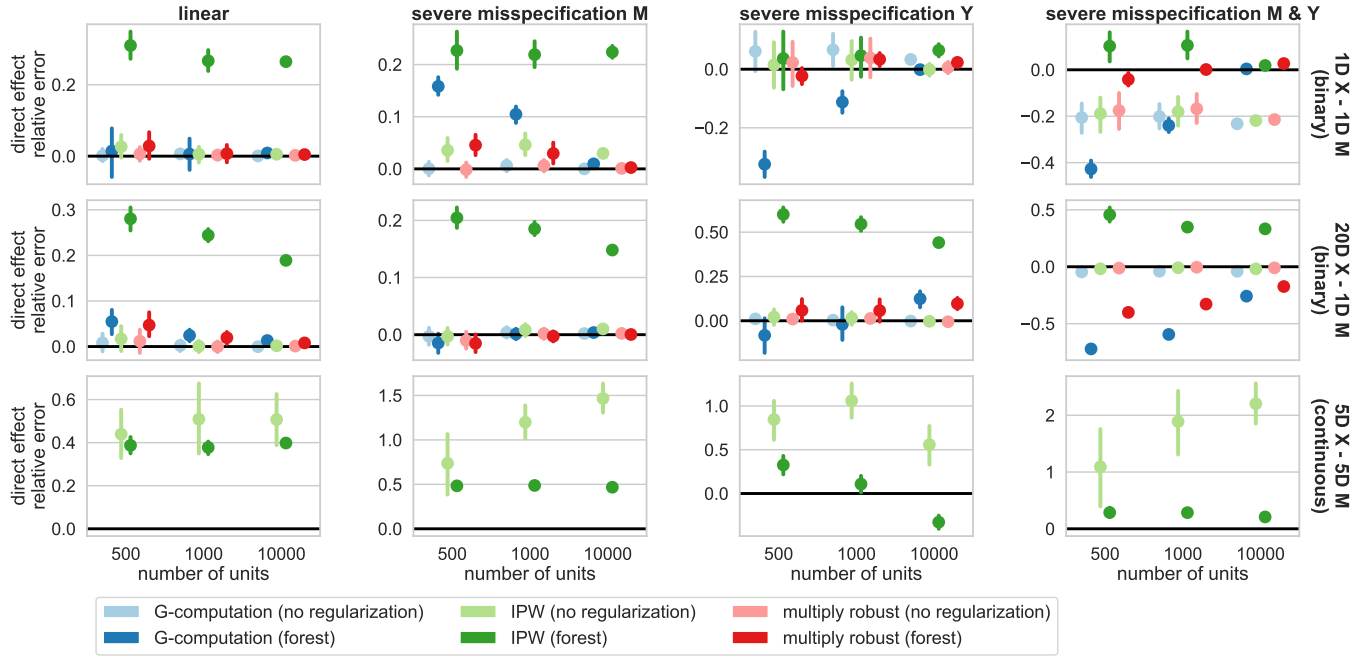

FIGURE 4 Effect of using a parametric or a non-parametric model for plug-in nuisance parameters estimation. We show results for four scenarii of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator and covariate dimensions, labeled on the right. Each dot represents the average relative error (i.e. $\frac{\hat{\theta}-\theta}{\theta}$ for the natural direct effect θ) over 30 repetitions, and the error bars are bootstrap 95% confidence intervals. All simulations are in the "medium" mediated proportion framework. The medDML method does not accept one-dimension covariates so we see no result for that estimator on the first row. Similarly, most estimators only handle binary one-dimension mediators, and have no results for the third row. We compare the estimators implemented in this study and compare linear parametric and random forest for fitting nuisance models. In most cases, the linear implementation has a smaller error than the random forest one, due to a smaller convergence rate, except in the case where both the mediator and the outcome models are misspecified.

6 | DISCUSSION AND CONCLUSION

In this study, we have conducted a thorough evaluation of the main estimators for mediation analysis, in a vast range of 378 distinct settings, covering in particular binary, continuous and multi-dimensional mediators, and several degrees of nuisance models misspecification. We also vary the number of observations, and the mediated proportions (strongly associated with violations of the positivity assumption for high indirect effect), which had not been considered for evaluation to the best of our knowledge. This allows us to assess the robustness of estimators to several violations of their consistency assumptions. Additionally, we have extended existing estimators to include recommended implementation strategies, namely regularization, use of non-parametric models, probability calibration and cross-fitting, with a total of 34 estimator variants. Our evaluation strategy also proposes a systematic analysis of their relevance.

Hence, this work allows us to provide practical recommendations. First of all, methods involving inverse probability weighting are unstable under some circumstances and should be used with great care, in particular regarding the overlap assumption, which is an issue in high-dimensional settings[47]. This nonetheless underlines that all other estimators are robust to violations of the overlap assumption. The use of random forests for nuisance parameter estimation requires more data, and hence generally hurts the performance at a fixed sample size, unless the parametric model is highly misspecified. The use of regularization and calibration has a very small effect, but tends to slightly improve performances. Finally cross-fitting is generally beneficial, except for small sample sizes, where it can increase the estimation error. Overall, the multiply robust estimator exhibits very promising results, even when non-parametric algorithms are used for nuisance parameters estimation. However, it only applies to binary mediators in its current implementation. In the multi-dimensional mediator setting, estimation seems unreliable, and the product of coefficients provides acceptable results in most settings.

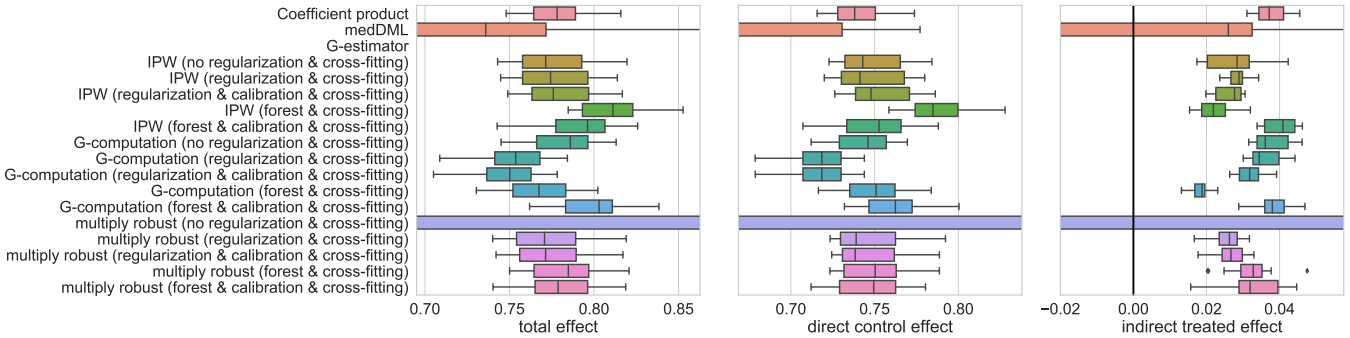

FIGURE 5 Mediation analysis of a physical job, for the effect on education on cognitive functions. The total, natural direct and indirect effects are shown in the panels from left to right. The scale was adjusted to show most results in details, so that the boxplots of the medDML and multiply robust (without regularization) estimators are cropped. The G-estimator method failed on this dataset. With the exception of medDML, G-estimator and the unregularized multiply robust estimator implementation, all estimators similar results for the total, direct and indirect effect, with a small but non-null indirect effect of a physical job.

An original finding of our study is the influence of the mediated proportion on the estimation error. A detailed analysis of simulated datasets exhibiting unstable predictions shows that a higher mediated proportion is associated with violations of the overlap assumption. As a consequence, the practitioner should consider the expected mediated proportion as an important criterion, and preferentially select an estimator robust to violations of the overlap assumption, in particular for multi-dimensional settings.

This article completes and extends a previous performance comparison of estimators for mediation analysis by Huber et. al[28]. First, we consider continuous and multi-dimensional mediators beyond the restrictive case of a binary mediators, although very few methods are currently available to handle them. We evaluate additional estimators or extensions of existing estimators that were not considered in this previous study. Finally, we also broaden the simulation spectrum to include violations of the linear model assumption, necessary for the consistency of some estimators, and highlight the sensitivity or robustness properties of the different estimators to violations of those assumptions.

Although the implementation of variants of existing estimators yields only mild improvements in terms of performance, it is of critical importance for off-the-shelf application to real data. Indeed, high-dimensional data are usually noisy, sometimes redundant, some variables might be uninformative due to a low variance in the sampled observations. In that setting, data cleaning is the most time-consuming step, so the robustness of implementations to corrupted data can highly impact the practitioner's ability to routinely use causal mediation analysis.

Our work presents however some limitations. The first and most important one lies on the use of simulated data, which is innately limited, despite our best efforts to cover a lot of different realistic settings. An aspect that was not considered in our simulations is the potential interactions between the treatment, the mediators and the covariates, which would require additional simulations.

This study highlights a number of potential research directions, in particular for estimators able to handle continuous or multi-dimensional mediators. As of now, only the coefficient product has satisfactory performances in most settings. A first line of work consists in increasing the stability of inverse-propensity based methods, with several solutions proposed outside of the mediation field[48,49]. Some clarification is also needed on how to best leverage more complex machine learning approaches to better estimate causal quantities.

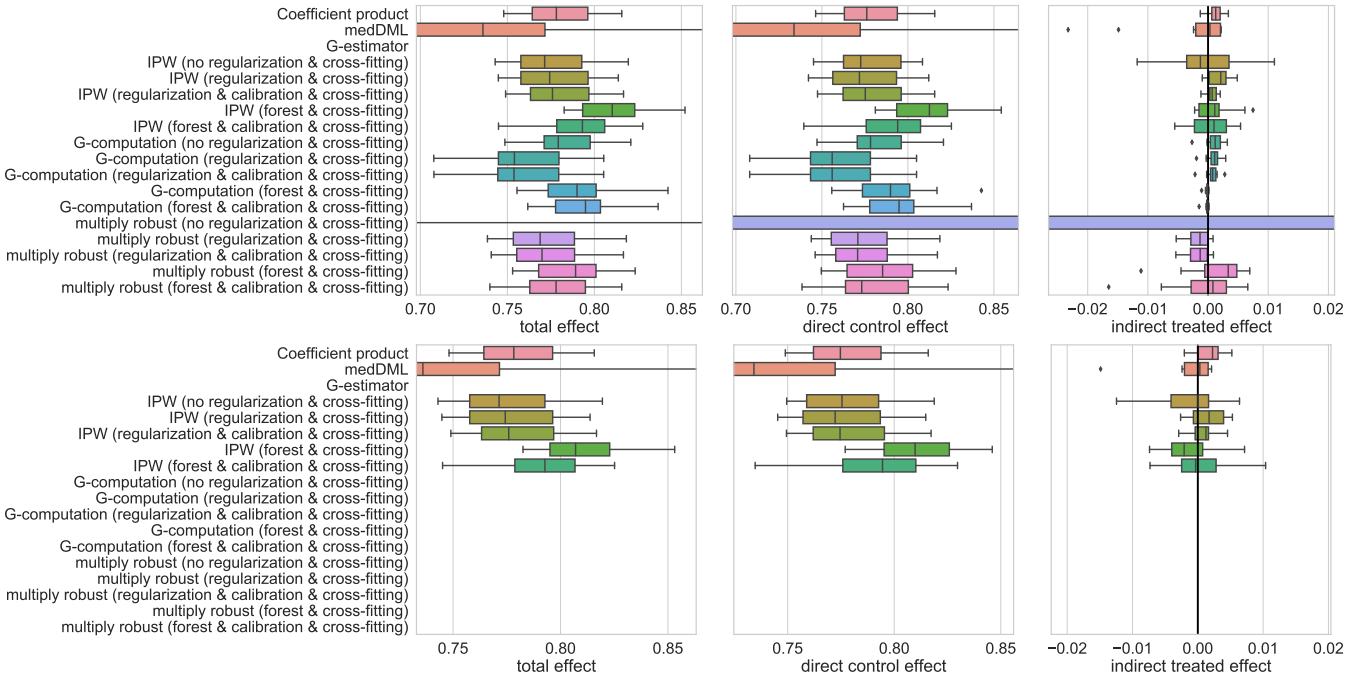

FIGURE 6 Mediation analysis of the brain age delta, binarized by its sign (panel a) or treated as a continuous variable (panel b). The total, natural direct and indirect effects are shown in the panels from left to right. The scale was adjusted to show most results in details, which crops the boxplots of the medDML and multiply robust (without regularization) estimators. For the binarized mediator (panel a), the G-estimator method failed on this dataset. In panel b, the considered mediator is continuous, so only the coefficient product, the IPW and the medDML estimators provided results. Both analyses consistently find no indirect effect through the brain age delta. With the exception of medDML, G-estimator and the unregularized multiply robust estimator implementation, all estimators found no indirect effect of the brain age delta, both for its binarized and continuous versions.

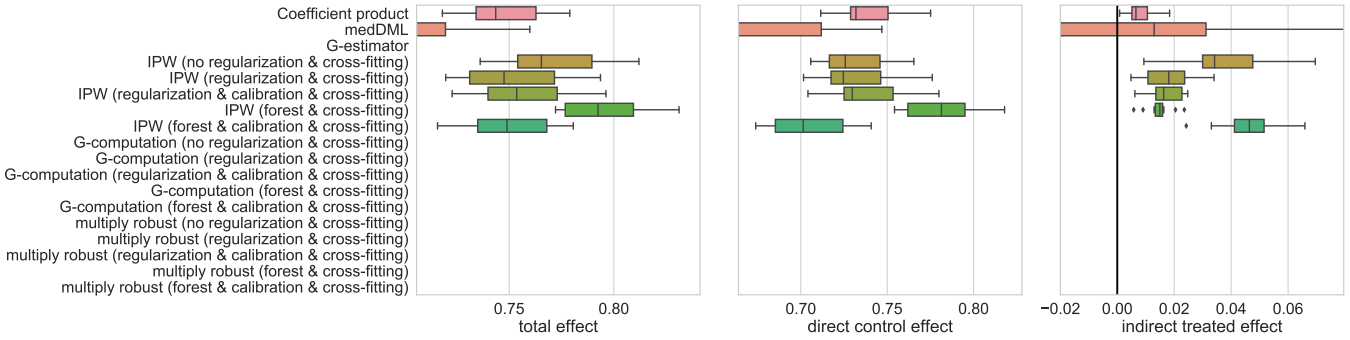

FIGURE 7 Mediation analysis of 10 variables summarizing brain imaging. The total, natural direct and indirect effects are shown in the panels from left to right. We adjusted the scale to show most results in details, which crops the boxplots of the medDML estimator. The mediator considered here has 10 dimensions, so only the coefficient product, the IPW and the medDML estimators provided results. All estimators but the medDML found a non-null indirect effect of the principal components extracted from the brain imaging variables.

References

1. Pearl J. Direct and indirect effects. In: ; 2001: 411–420.
2. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992: 143–155.
3. Imbens GW, Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press . 2015.
4. Cochran WG. Analysis of covariance: its nature and uses. *Biometrics* 1957; 13(3): 261–281.
5. Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Evaluation review* 1981; 5(5): 602–619.
6. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.. *Journal of personality and social psychology* 1986; 51(6): 1173.
7. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. In: Green PJ, Richardson S, Hjort ., eds. *In Highly Structured Stochastic Systems*Oxford University Press. 2003 (pp. 70–81).
8. Petersen ML, Sinisi SE, Laan v. dMJ. Estimation of direct causal effects. *Epidemiology* 2006: 276–284.
9. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science* 2010; 25(1): 51–71.
10. Hong G. Ratio of mediator probability weighting for estimating natural direct and indirect effects. In: American Statistical Association Alexandria, VA. ; 2010: 2401–2415.
11. Albert JM, Nelson S. Generalized causal mediation analysis. *Biometrics* 2011; 67(3): 1028–1038.
12. Tchetgen EJT, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics* 2012; 40(3): 1816.
13. Huber M. Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics* 2014; 29(6): 920–943.
14. Zheng W, Laan v. dMJ. Targeted maximum likelihood estimation of natural direct effects. *The international journal of biostatistics* 2012; 8(1): 1–40.
15. Farbmacher H, Huber M, Lafférs L, Langen H, Spindler M. Causal mediation analysis with double machine learning. *arXiv preprint arXiv:2002.12710* 2020.
16. Hines O, Vansteelandt S, Diaz-Ordaz K. Robust inference for mediated effects in partially linear models. *Psychometrika* 2021: 1–24.
17. VanderWeele T, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiologic methods* 2014; 2(1): 95–115.
18. Jérolon A, Baglietto L, Birmelé E, Alarcon F, Perduca V. Causal mediation analysis in presence of multiple mediators uncausally related. *The International Journal of Biostatistics* 2020.
19. Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. 2005.
20. Huber M. A review of causal mediation analysis for assessing direct and indirect treatment effects. 2019.
21. Nguyen TQ, Schmid I, Ogburn EL, Stuart EA. Clarifying causal mediation analysis for the applied researcher: Effect identification via three assumptions and five potential outcomes. *arXiv preprint arXiv:2011.09537* 2020.
22. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA. High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* 2018; 19(2): 121–136. Publisher: Oxford Academicdoi: 10.1093/biostatistics/kxx027

23. Zhao Y, Lindquist MA, Caffo BS. Sparse principal component based high-dimensional mediation analysis. *Computational Statistics and Data Analysis* 2020; 142: 106835.

24. Huang YT, Pan WC. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* 2016; 72(2): 402–413.

25. Zhang H, Chen J, Feng Y, Wang C, Li H, Liu L. Mediation effect selection in high-dimensional and compositional microbiome data. *Statistics in medicine* 2021; 40(4): 885–896.

26. Djordjilović V, Page CM, Gran JM, et al. Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in medicine* 2019; 38(18): 3346–3360.

27. MacKinnon DP. *Introduction to statistical mediation analysis*. Routledge . 2012.

28. Huber M, Lechner M, Mellace G. The finite sample performance of estimators for mediation analysis under sequential conditional independence. *Journal of Business & Economic Statistics* 2016; 34(1): 139–160.

29. VanderWeele T. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press . 2015.

30. Scott DW, Sain SR. Multi-dimensional Density Estimation. 2004.

31. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis.. *Psychological methods* 2010; 15(4): 309.

32. Vermeulen K, Vansteelandt S. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* 2015; 110(511): 1024–1036.

33. Avagyan V, Vansteelandt S. Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation. *arXiv preprint arXiv:1708.03787* 2017.

34. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. 2018.

35. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; 12: 2825–2830.

36. Breiman L. Random forests. *Machine learning* 2001; 45(1): 5–32.

37. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: ; 2002: 694–699.

38. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: ; 2005: 625–632.

39. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. Mediation: R package for causal mediation analysis. 2014.

40. Bodory H, Huber M. The causalweight package for causal inference in R. tech. rep., University of Fribourg; 2018.

41. Nguyen TQ, Schmid I, Stuart EA. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn.. *Psychological Methods* 2021; 26(2): 255.

42. Ritchie SJ, Tucker-Drob EM. How much does education improve intelligence? A meta-analysis. *Psychological science* 2018; 29(8): 1358–1369.

43. Whalley LJ, Deary IJ, Appleton CL, Starr JM. Cognitive reserve and the neurobiology of cognitive aging. *Ageing research reviews* 2004; 3(4): 369–382.

44. Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience* 2016; 19(11): 1523–1536.

45. Fawns-Ritchie C, Deary IJ. Reliability and validity of the UK Biobank cognitive tests. *PloS one* 2020; 15(4): e0231627.

46. Nyberg L, Magnussen F, Lundquist A, et al. Educational attainment does not influence brain aging. *Proceedings of the National Academy of Sciences* 2021; 118(18).

47. D'Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics* 2021; 221(2): 644–654.

48. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology* 2019; 188(1): 250–257.

49. Conover MM, Rothman KJ, Stürmer T, Ellis AR, Poole C, Jonsson Funk M. Propensity score trimming mitigates bias due to covariate measurement error in inverse probability of treatment weighted analyses: A plasmode simulation. *Statistics in Medicine* 2021; 40(9): 2101–2112.

How to cite this article: Abécassis J., Josse J., and Thirion B. (2022), Evaluation of mediation analysis methods in a variety of contexts, , .

APPENDIX

S1 IDENTIFIABILITY OF THE TOTAL EFFECT, AND THE NATURAL DIRECT AND INDIRECT EFFECT

In this section, we demonstrate that the potential outcomes defined in Section 2 are identified under Assumptions 1, 2, 3 and 4. We denote $f_{A=a}$ and $f_{A=a|B=b}$ the probability density function of a random variable A, unconditionally or conditionally given other random variable(s) B:

Let us first consider the potential outcomes $Y(t, M(t))$:

$$
\begin{aligned}
&\mathbb{E}\left[Y\left(t, M\left(t\right)\right)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[Y(t, M(t))|X = x\right]\right] && \text{sequential ignorability (assumption ??)} \\
=&\mathbb{E}\left[\mathbb{E}\left[Y(t, M(t))|T = t, X = x\right]\right] && \text{SUTVA (consistency)} \\
=&\mathbb{E}\left[\mathbb{E}\left[Y|T = t, X = x\right]\right] && \text{overlap (assumption ??)} \\
=&\mathbb{E}\left[\mathbb{E}\left[\frac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|X)}|X = x\right]\right] \\
=&\mathbb{E}\left[\frac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|X)}\right] && \text{(A1)}
\end{aligned}
$$

And then the cross-world potential outcomes $Y(t, M(t'))$.

$$
\begin{aligned}
&\mathbb{E}\left[Y\left(t, M\left(t'\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y\left(t, M\left(t'\right)\right)|X = x\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y\left(t, m\right)\right)|X = x, M(t') = m\right]\right] \\
=&\int\int \mathbb{E}\left[Y(t, m)|M\left(t'\right) = m, X = x\right] f_{M(t')=m|X=x} dm f_{X=x} dx && \text{ignorability (assumption ??)} \\
=&\int\int \mathbb{E}[Y|T = t, M = m, X = x] f_{M(t')=m|X=x} f_{X=x} dm dx && \text{and consistency (SUTVA)} \\
=&\int\int \mathbb{E}[Y|T = t, M = m, X = x] f_{M=m|T=t',X=x} f_{X=x} dm dx && \text{ignorability (assumption ??)} \quad \text{(A2)} \\
=&\int\int \mathbb{E}[Y|T = t, M = m, X = x] \cdot \frac{\mathbb{P}\left(T=t'|M=m,X=x\right)}{\mathbb{P}(T=t'|X=x)} f_{M=m|X=x} dm f_{X=x} dx \\
=&\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\frac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|M,X)}|M, X\right] \cdot \frac{\mathbb{P}\left(T=t'|M,X\right)}{\mathbb{P}(T=t'|X)}|X\right]\right] \\
=&\mathbb{E}\left[\frac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|M,X)} \cdot \frac{\mathbb{P}\left(T=t'|M,X\right)}{\mathbb{P}(T=t'|X)}\right] && \text{(A3)} \\
=&\mathbb{E}\left[\frac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|X)} \cdot \frac{f\left(M=m|T=t',X\right)}{f(M=m|T=t,X)}\right] && \text{(A4)}
\end{aligned}
$$

S2 EXTENDED RESULTS ON SIMULATIONS

In this section, we report more thoroughly the results on simulations, including results for the total and indirect effects, and the effect of calibration and cross-fitting for all nuisance parameters estimation algorithms.

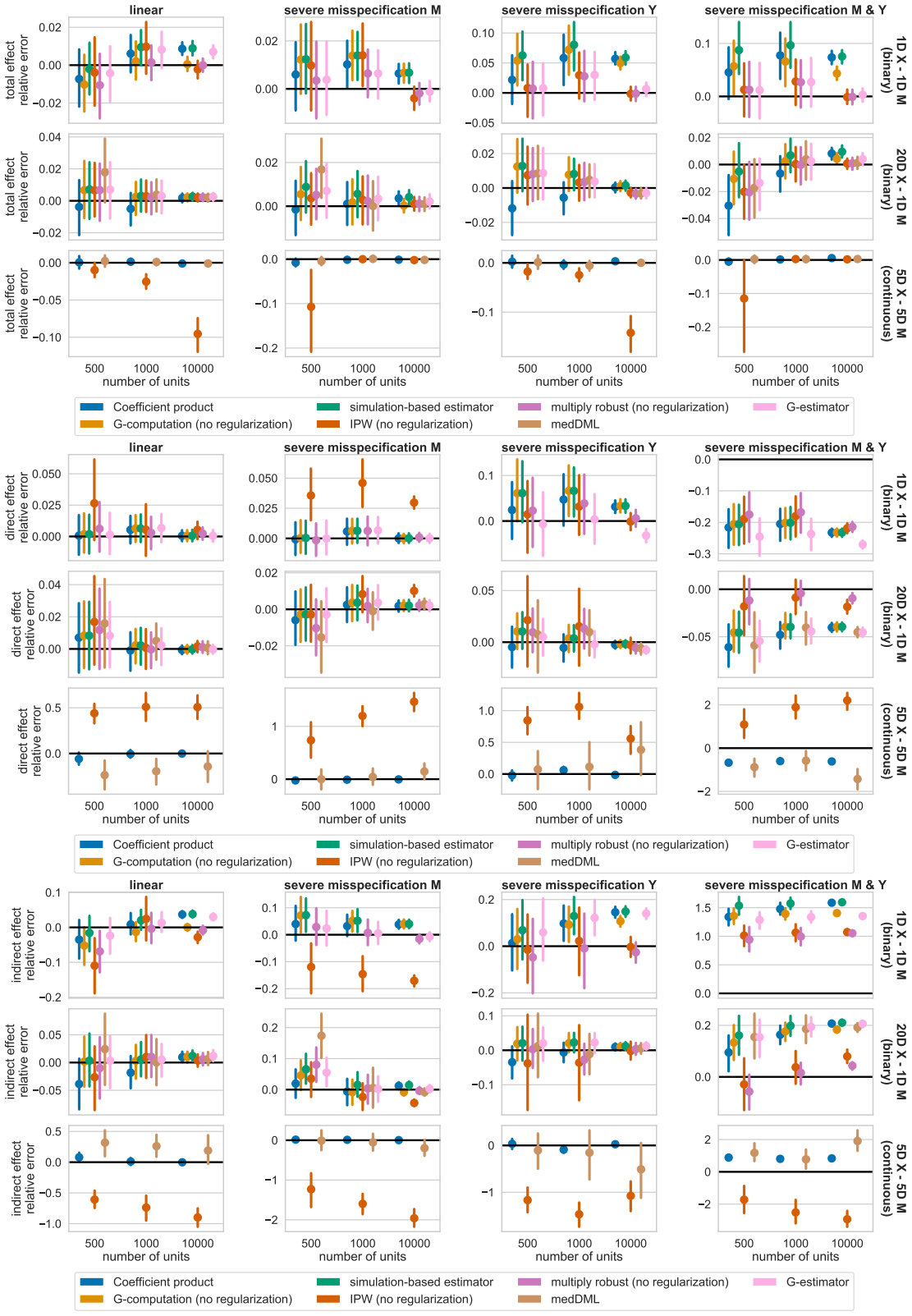

FIGURE S1 Effect of sample size on total, direct and indirect effects estimation. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

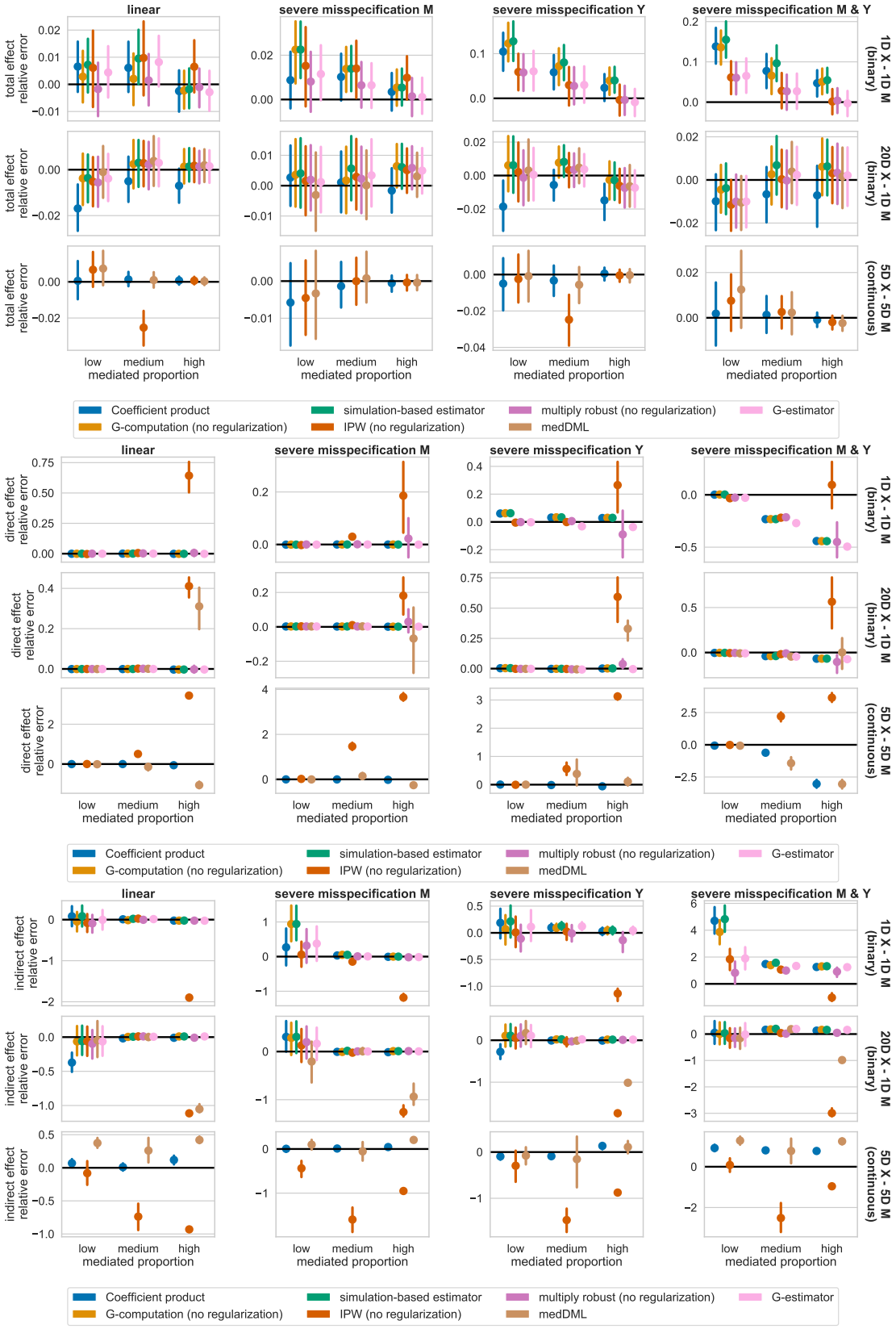

FIGURE S2 Effect of the mediated proportion on total, direct and indirect effects estimation. All simulations are with $n = 1000$ observations. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

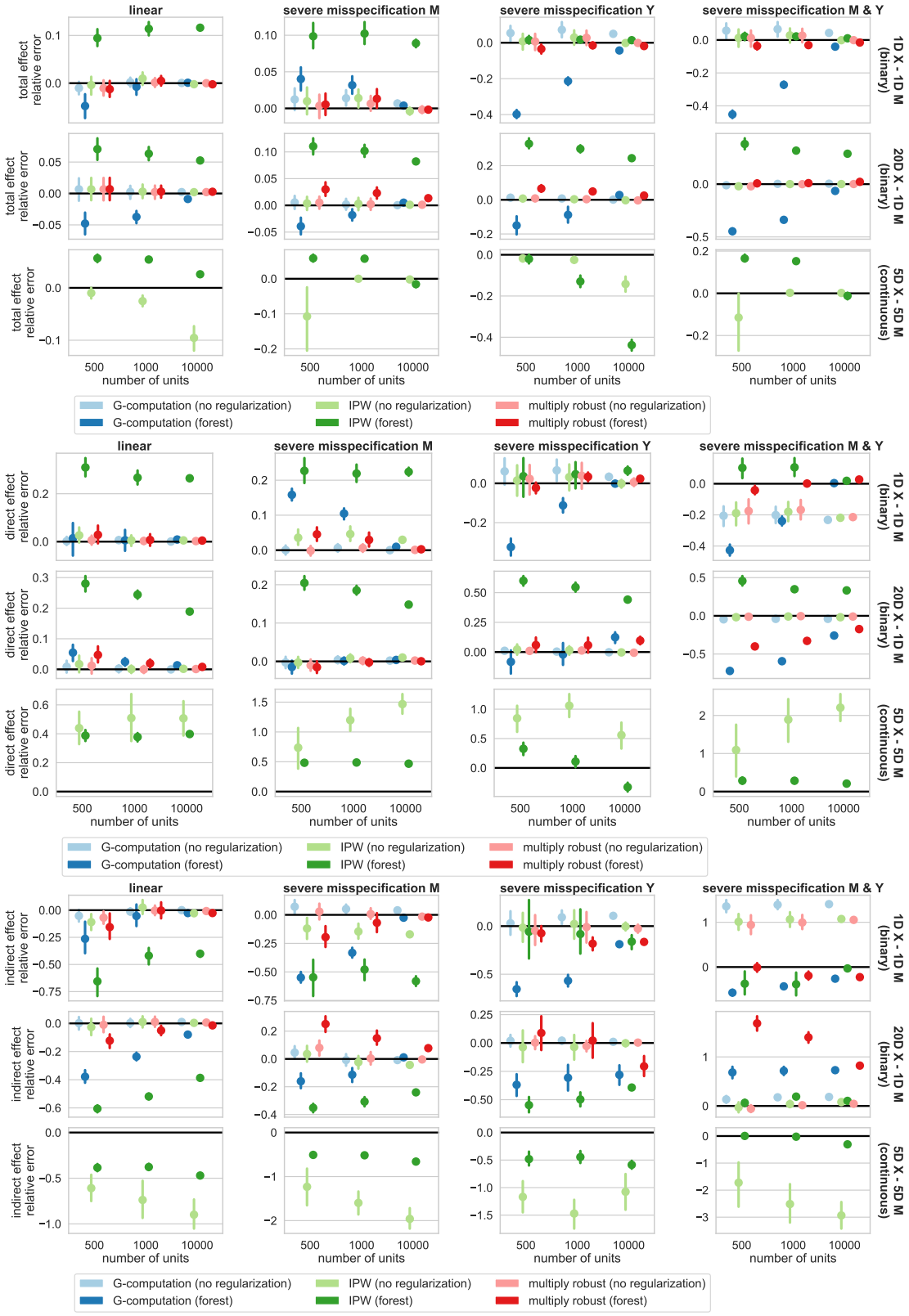

FIGURE S3 Effect of using a parametric or a non-parametric model for plug-in nuisance parameters estimation. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

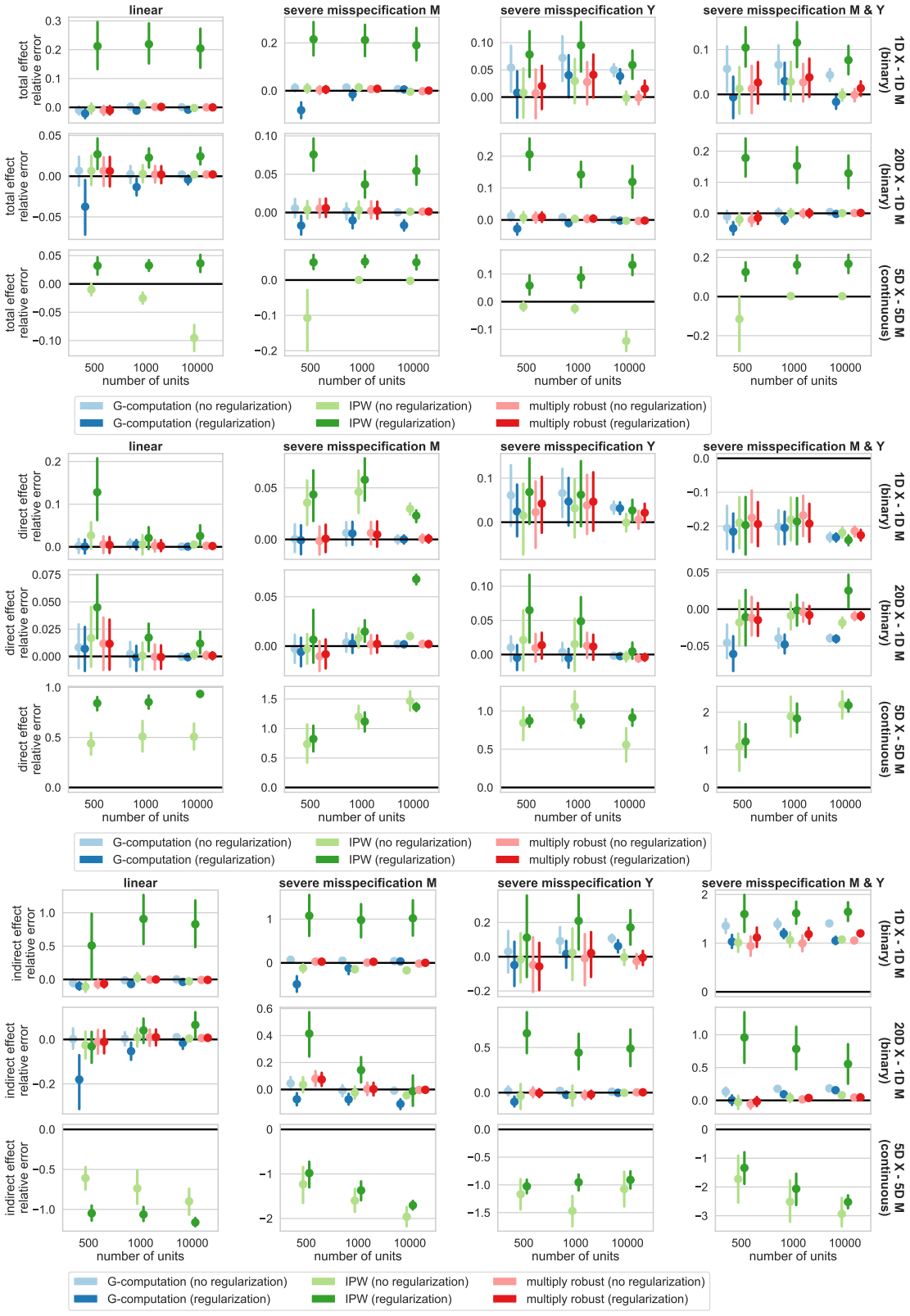

FIGURE S4 Effect of using regularization for plug-in nuisance parameters estimation. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

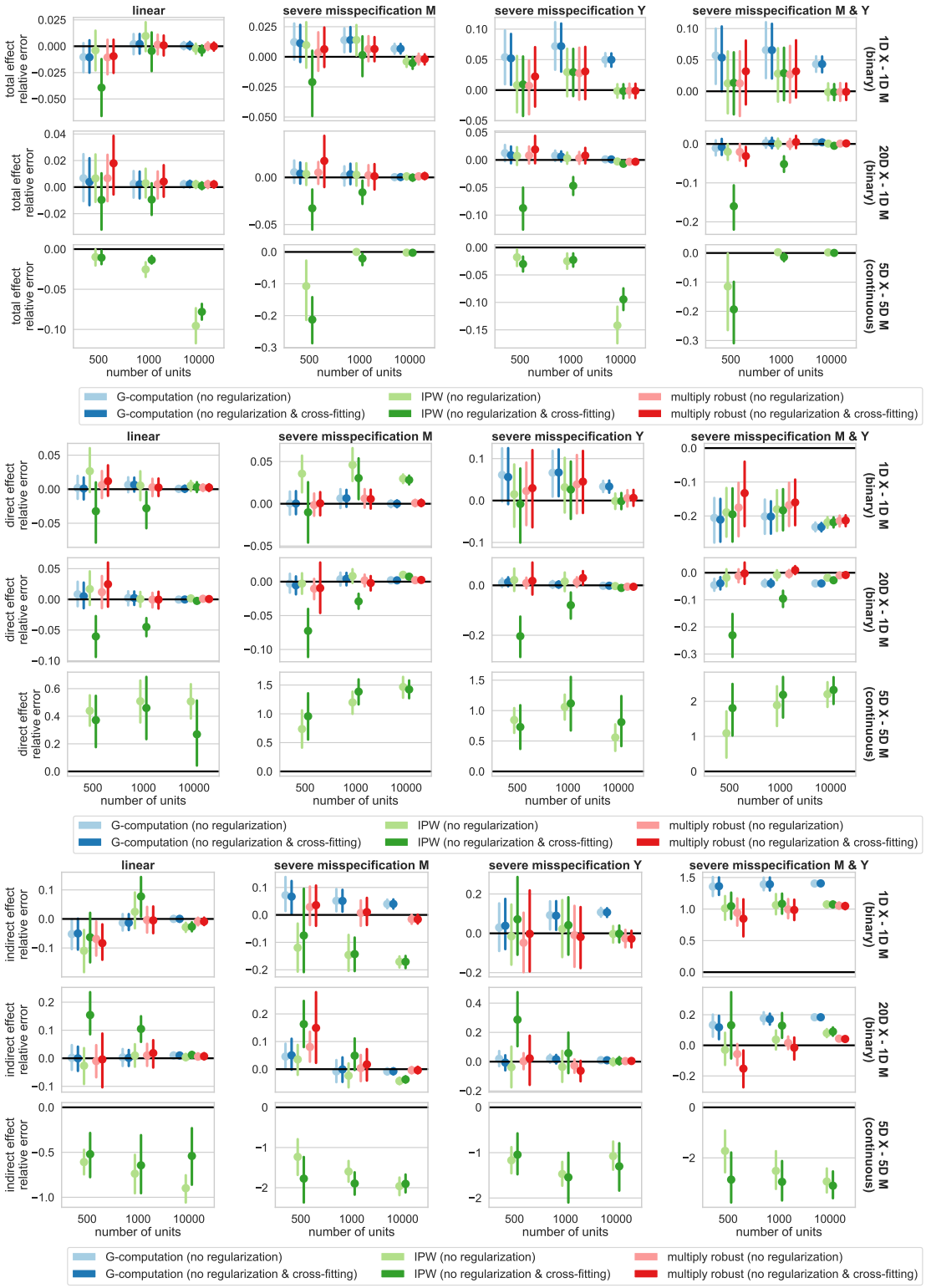

FIGURE S5 Effect of using cross-fitting for plug-in nuisance parameters estimation with a non-regularized parametric model. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

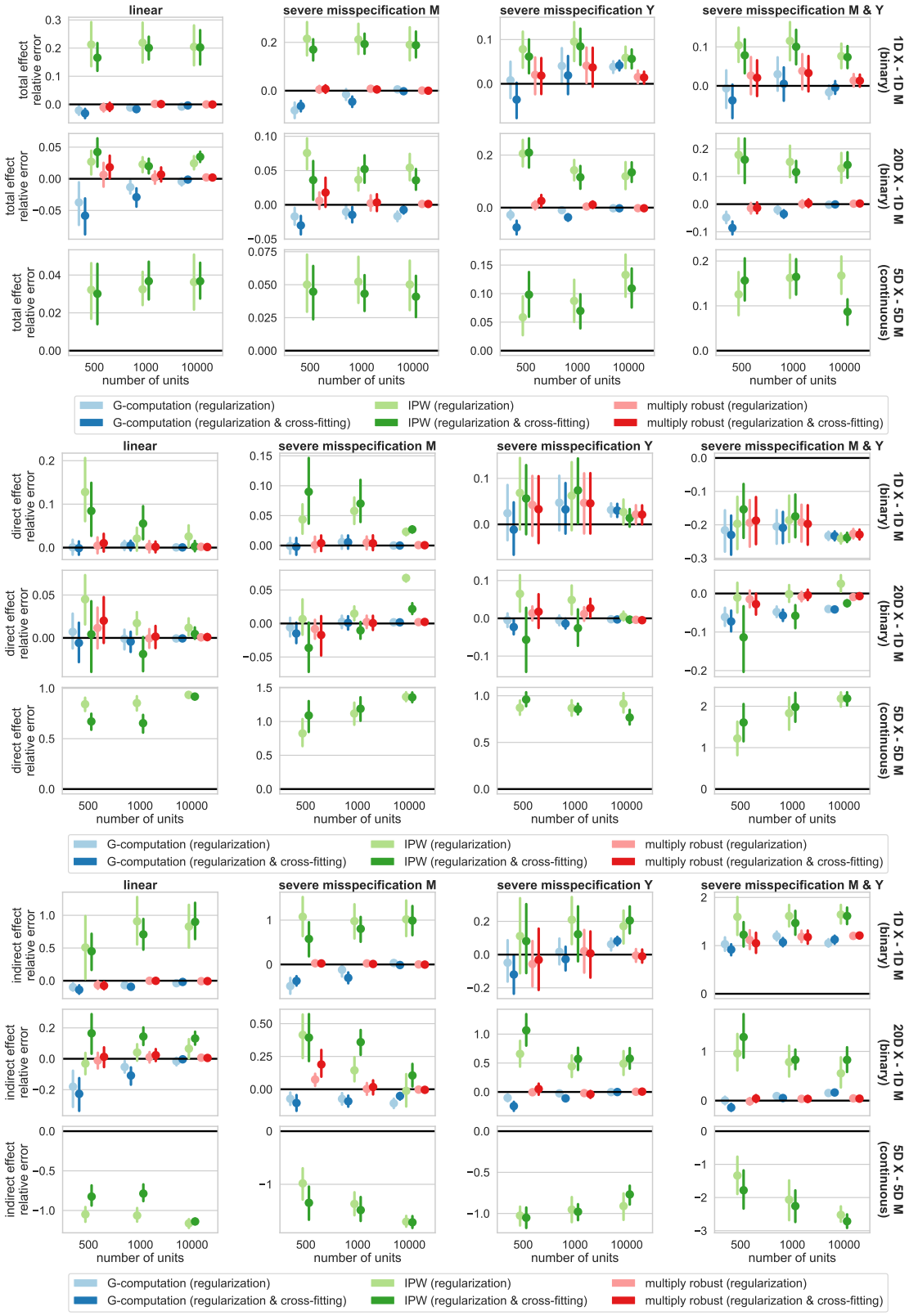

FIGURE S6 Effect of using cross-fitting for plug-in nuisance parameters estimation with a regularized parametric model. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

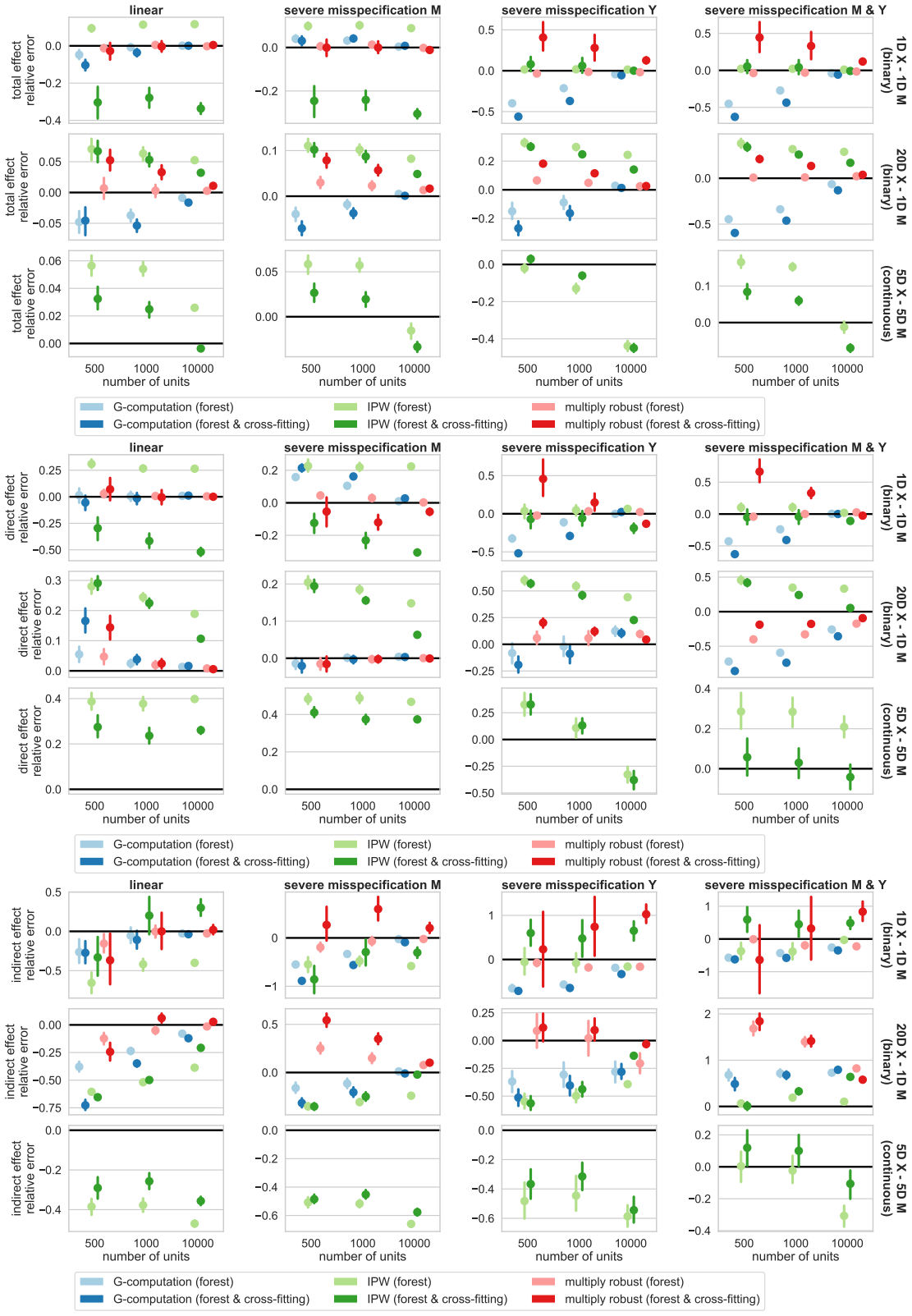

FIGURE S7 Effect of using cross-fitting for plug-in nuisance parameters estimation with a non-parametric model. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

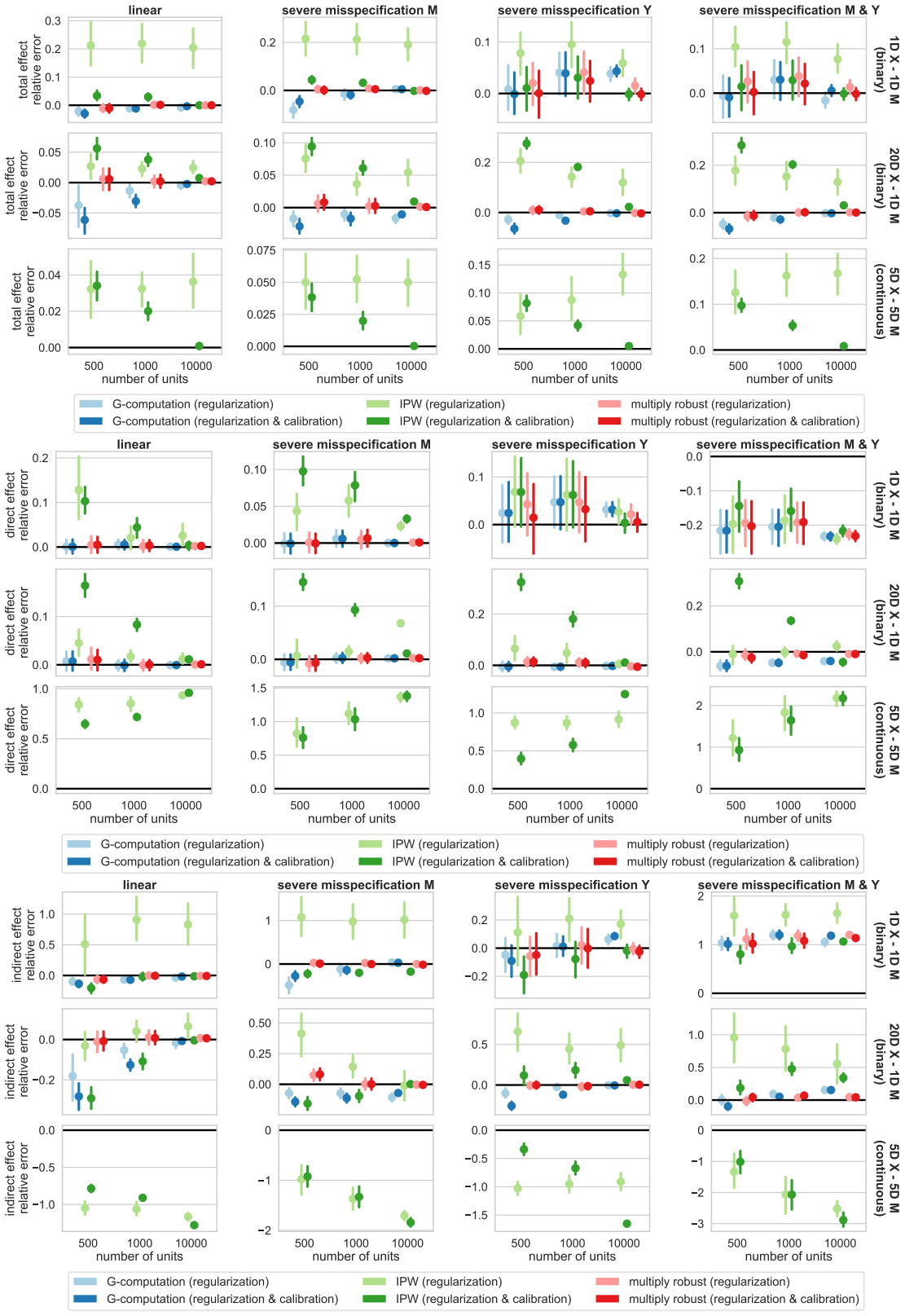

FIGURE S8 Effect of using calibration for plug-in nuisance parameters estimation with a parametric regularized model. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

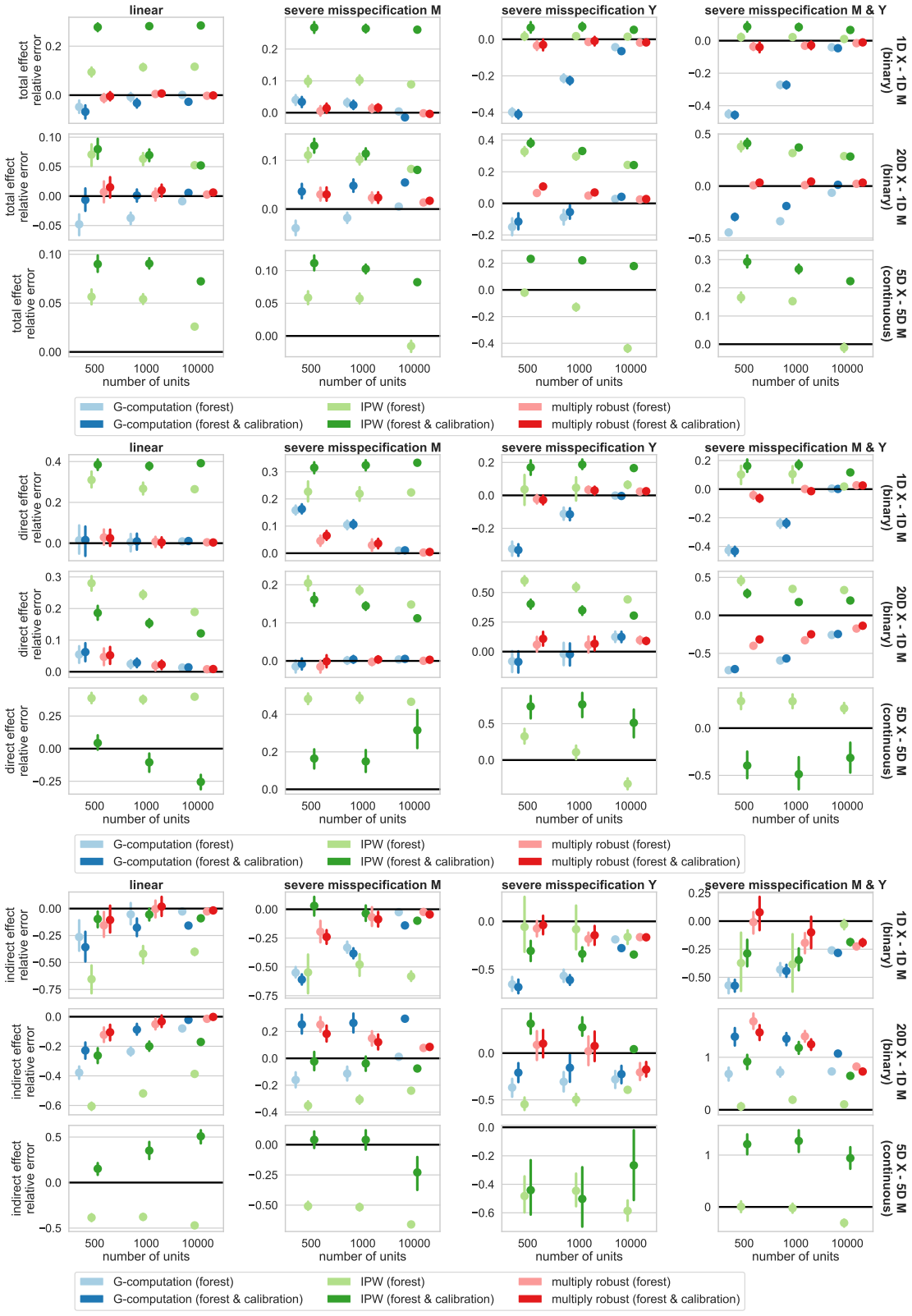

FIGURE S9 Effect of using calibration for plug-in nuisance parameters estimation with a non-parametric model. All simulations are in the "medium" mediated proportion framework. Four scenarii (in columns) and three mediator and covariates dimensions (in rows) are considered. The mediated proportion was fixed to "medium".

S3 UK BIOBANK APPLICATION

To apply mediation analysis on the cognitive functions in the UK Biobank, we have considered four potential mediators. For each of them, we have selected relevant confounding variables. We have selected participants for which all required variables were available for the four problems, resulting in a subset of 16,157 participants. The list of all used variables is presented in Supplementary Table S1.

Variable role	constructed summary	**Field or category ID**	**Variable name**
outcome	g-factor	Field 398	Number of correct matches in round
		Field 20023	Mean time to correctly identify matches
		Field 20018	Prospective memory result
		Field 20016	Fluid intelligence score
		Field 6348	Duration to complete numeric path
		Field 6350	Duration to complete alphanumeric path
		Field 23324	Number of symbol digit matches made correctly
		Field 21004	Number of puzzles correct
		Field 6373	Number of puzzles correctly solved
		Field 20197	Number of word pairs correctly associated
mediator	10 PCs	Category 134	Diffusion MRI skeleton measurements
		Category 1101	Regional grey matter volumes (FAST)
		Field 25001	Volume of peripheral cortical grey matter (normalised for head size)
		Field 25005	Volume of grey matter (normalised)
		Field 25009	Volume of brain, grey+white matter (normalised)
		Field 25011	Volume of thalamus (left)
		Field 25012	Volume of thalamus (right)
		Field 25013	Volume of caudate (left)
		Field 25014	Volume of caudate (right)
		Field 25015	Volume of putamen (left)
		Field 25016	Volume of putamen (right)
		Field 25017	Volume of pallidum (left)
		Field 25018	Volume of pallidum (right)
		Field 25019	Volume of hippocampus (left)
		Field 25020	Volume of hippocampus (right)
		Field 25021	Volume of amygdala (left)
		Field 25022	Volume of amygdala (right)
		Field 25023	Volume of accumbens (left)
		Field 25024	Volume of accumbens (right)
		Field 25025	Volume of brain stem + 4th ventricle)
mediator	at least one of	Field 806	Job involves mainly walking or standing
		Field 816	Job involves heavy manual or physical work
		Field 826	Job involves shift work
confounders		Field 31	Sex
		Field 21003	Age when attended assessment centre
		Field 54	UK Biobank assessment centre
		Field 6142	Current employment status
		Field 189	Townsend deprivation index at recruitment
		Field 20160	Ever smoked
		Field 1239	Current tobacco smoking
		Field 1249	Past tobacco smoking
		Field 1647	Country of birth (UK/elsewhere)
		Field 1677	Breastfed as a baby
		Field 1687	Comparative body size at age 10
		Field 1697	Comparative height size at age 10
		Field 1707	Handedness (chirality/laterality)
		Field 1767	Adopted as a child
		Field 1777	Part of a multiple birth
		Field 1787	Maternal smoking around birth
		Field 1558	Alcohol intake frequency.
		Field 21001	Body mass index (BMI)
		Field 25000	Volumetric scaling from T1 head image to standard space
		Field 25756	Scanner lateral (X) brain position
		Field 25757	Scanner transverse (Y) brain position
		Field 25758	Scanner longitudinal (Z) brain position
		Field 25759	Scanner table position
treatment		Field 6138	Qualifications

TABLE S1 UK Biobank Field codes for all variables used