S1 Identifiability of the total effect, and the natural direct and indirect effect

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

$$
\begin{aligned}
&\mathbb{E}\left[Y\left(t, M\left(t\right)\right)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[Y(t, M(t))|X=x\right]\right] && \text{sequential ignorability (assumption 2)} \\
=&\mathbb{E}\left[\mathbb{E}\left[Y(t, M(t))|T=t, X=x\right]\right] && \text{SUTVA (assumption 1)} \\
=&\mathbb{E}\left[\mathbb{E}\left[Y|T=t, X=x\right]\right] && \text{overlap (assumption 4)} \\
=&\mathbb{E}\left[\mathbb{E}\left[\tfrac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|X)}|X=x\right]\right] \\
=&\mathbb{E}\left[\tfrac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|X)}\right] && (16)
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathbb{E}\left[Y\left(t, M\left(t'\right)\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y\left(t, M\left(t'\right)\right)|X=x\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y\left(t, m\right)|X=x, M(t')=m\right]\right]\right] \\
=& \iint \mathbb{E}\left[Y(t,m)|M\left(t'\right)=m, X=x\right] f_{M(t')=m|X=x} dm f_{X=x} dx && \text{ignorability (assumption 2)} \\
=& \iint \mathbb{E}[Y|T=t, M=m, X=x] f_{M(t')=m|X=x} f_{X=x} dm dx && \text{and SUTVA (assumption 1)} \\
=& \iint \mathbb{E}[Y|T=t, M=m, X=x] f_{M=m|T=t', X=x} f_{X=x} dm dx && \text{ignorability (assumption 3)} \\
& && (17) \\
=& \iint \mathbb{E}[Y|T=t, M=m, X=x] \cdot \tfrac{\mathbb{P}\left(T=t'|M=m, X=x\right)}{\mathbb{P}(T=t'|X=x)} f_{M=m|X=x} dm f_{X=x} dx \\
=& \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\tfrac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|M,X)}|M, X\right] \cdot \tfrac{\mathbb{P}\left(T=t'|M,X\right)}{\mathbb{P}(T=t'|X)}|X\right]\right] \\
=& \mathbb{E}\left[\tfrac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|M,X)} \cdot \tfrac{\mathbb{P}\left(T=t'|M,X\right)}{\mathbb{P}(T=t'|X)}\right] && (18) \\
=& \mathbb{E}\left[\tfrac{Y \cdot I\{T=t\}}{\mathbb{P}(T=t|X)} \cdot \tfrac{f\left(M=m|T=t', X\right)}{f(M=m|T=t, X)}\right] && (19)
\end{aligned}
$$

S2 More details about the causal mediation estimators derivation

S2.1 G-computation without mediator density estimation

We recall the mediation formula

$$\theta(t) = \iint \{\mathbb{E}[Y \mid T=1, M=m, X=x] - \mathbb{E}[Y \mid T=0, M=m, X=x]\} f_{M=m|T=t,X=x} dm f_{X=x} dx$$

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

Instead of doing the double integration explicitly in the estimation, we can integrate over the mediator in the mediation formula above, yielding:

$$\theta(t) = \int \left(\mathbb{E}\left[\mu_Y(1, m, x)|T=t, X=x\right] - \mathbb{E}\left[\mu_Y(0, m, x)|T=t, X=x\right]\right) f_{X=x} dx$$

$$\delta(t) = \int \left(\mathbb{E}\left[\mu_Y(t, m, x)|T=1, X=x\right] - \mathbb{E}\left[\mu_Y(t, m, x)|T=0, X=x\right]\right) f_{X=x} dx.$$

with $\mu_Y(t, m, x) = \mathbb{E}[Y \mid T=t, M=m, X=x]$. Those terms involving two different values of the treatment are called cross-world conditional mean outcomes, $\omega_Y(t, t', x) = \mathbb{E}\left[\mu_Y(t, m, x)|T=t', X=x\right]$, following the notations of [Zenati et al., 2025], and represent the expected outcome for values of the treatment t in the population that has actually received the treatment t'. It can be estimated by

1. fit a regression model to predict Y from (T, M, X) (model $\hat{\mu}$)
2. fit a second regression model on the predictions $\hat{Y}$ from $\hat{\mu}$ on X, in the population where $T = t'$ (model $\hat{\nu}_{t'}(\hat{\mu}(t, m, x), x)$)

3. take the empirical mean of those predictions from the second regression over the whole sample

This leads to the following for estimation: $\hat{\omega_Y}(t, t', x) = \hat{\nu}_{t'}(\hat{\mu}(t, M_i, X_i), X_i)$

$$
\begin{aligned}
\hat{\theta}(t) &= \frac{1}{n}\sum_{i=1}^{n} \hat{\omega_Y}(1, t, x) - \hat{\omega_Y}(0, t, x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{\nu}_t(\hat{\mu}(1, M_i, X_i), X_i) - \hat{\nu}_t(\hat{\mu}(0, M_i, X_i), X_i) \\
\hat{\delta}(t) &= \frac{1}{n}\sum_{i=1}^{n} \hat{\omega_Y}(t, 1, x) - \hat{\omega_Y}(t, 0, x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{\nu}_1(\hat{\mu}(t, M_i, X_i), X_i) - \hat{\nu}_0(\hat{\mu}(t, M_i, X_i), X_i)
\end{aligned}
\tag{20}
$$

This new formula does not necessitate the estimation of the mediator density, but two nested regression steps.

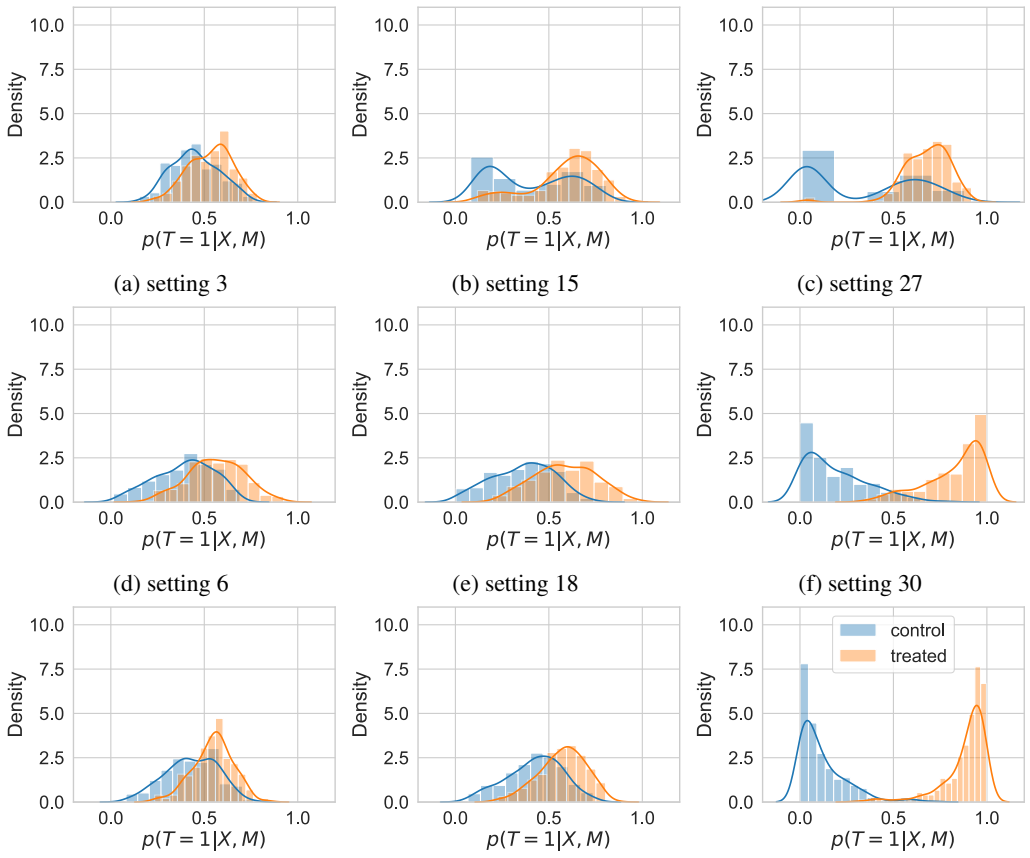
S3 More details on simulation settings

We describe the main parameters of each considered simulation setting in Table S1.

setting_nb	n	dim_x	dim_m	type_m	wt_list	wm_list	m_misspec	y_misspec	total	direct_1	direct_0	indirect_1	indirect_0
1	500	5	1	binary	0.50	0.50	False	False	1.23	1.20	1.20	0.03	0.03
2	500	5	1	binary	0.50	0.50	True	False	1.23	1.20	1.20	0.03	0.03
3	500	5	1	binary	0.50	0.50	False	True	1.54	1.51	1.45	0.09	0.03
4	500	5	1	binary	0.50	0.50	True	True	1.53	1.51	1.45	0.08	0.03
5	500	5	1	continuous	0.50	0.50	False	False	1.33	1.20	1.20	0.13	0.13
6	500	5	1	continuous	0.50	0.50	True	False	1.32	1.20	1.20	0.12	0.12
7	500	5	1	continuous	0.50	0.50	False	True	1.58	1.45	1.20	0.38	0.13
8	500	5	1	continuous	0.50	0.50	True	True	1.60	1.46	1.20	0.40	0.13
9	500	5	5	continuous	0.20	0.20	False	False	1.22	1.20	1.20	0.02	0.02
10	500	5	5	continuous	0.20	0.20	True	False	1.22	1.20	1.20	0.02	0.02
11	500	5	5	continuous	0.20	0.20	False	True	1.31	1.29	1.19	0.11	0.02
12	500	5	5	continuous	0.20	0.20	True	True	1.32	1.30	1.20	0.11	0.02
13	500	5	1	binary	2.00	10.00	False	False	3.05	1.20	1.20	1.85	1.85
14	500	5	1	binary	2.00	10.00	True	False	3.00	1.20	1.20	1.80	1.80
15	500	5	1	binary	2.00	10.00	False	True	3.52	1.64	1.45	2.07	1.88
16	500	5	1	binary	2.00	10.00	True	True	3.43	1.63	1.45	1.98	1.80
17	500	5	1	continuous	0.80	5.00	False	False	3.27	1.20	1.20	2.07	2.07
18	500	5	1	continuous	0.80	5.00	True	False	2.95	1.20	1.20	1.75	1.75
19	500	5	1	continuous	0.80	5.00	False	True	3.24	1.57	1.17	2.01	1.67
20	500	5	1	continuous	0.80	5.00	True	True	3.75	1.61	1.21	2.57	2.14
21	500	5	5	continuous	0.30	5.00	False	False	1.90	1.20	1.20	0.70	0.70
22	500	5	5	continuous	0.30	5.00	True	False	1.97	1.20	1.20	0.77	0.77
23	500	5	5	continuous	0.30	5.00	False	True	2.03	1.34	1.19	0.82	0.69
24	500	5	5	continuous	0.30	5.00	True	True	2.13	1.35	1.20	0.94	0.78
25	500	5	1	binary	4.00	2.00	False	False	1.67	1.20	1.20	0.47	0.47
26	500	5	1	binary	4.00	2.00	True	False	1.68	1.20	1.20	0.48	0.48
27	500	5	1	binary	4.00	2.00	False	True	2.17	1.69	1.45	0.72	0.48
28	500	5	1	binary	4.00	2.00	True	True	2.16	1.69	1.45	0.71	0.47
29	500	5	1	continuous	2.00	1.00	False	False	2.26	1.20	1.20	1.06	1.06
30	500	5	1	continuous	2.00	1.00	True	False	2.18	1.20	1.20	0.98	0.98
31	500	5	1	continuous	2.00	1.00	False	True	3.20	2.20	1.20	2.01	1.00
32	500	5	1	continuous	2.00	1.00	True	True	3.20	2.20	1.20	2.00	1.00
33	500	5	5	continuous	1.00	1.00	False	False	1.69	1.20	1.20	0.49	0.49
34	500	5	5	continuous	1.00	1.00	True	False	1.71	1.20	1.20	0.51	0.51
35	500	5	5	continuous	1.00	1.00	False	True	2.22	1.71	1.21	1.02	0.51
36	500	5	5	continuous	1.00	1.00	True	True	2.20	1.70	1.20	1.00	0.50

Table S1: **parameter settings for simulations.** This table presents the different sets of parameters chosen for the simulations. The different color blocks represent modulations of the indirect effect size, with a low indirect effect size (green), a high indirect effect size but no overlap violation (orange) or a high indirect effect size with a strong overlap violation (gray). Those effect sizes are adjusted either through ω_M which sets the strength of the relation between M and Y, or through ω_T which sets the strength of the relation between T and M. A value of ω_T too high can lead to the violation of the overlap assumption.

In Figure S1, we illustrate the respect of the overlap assumption 4 by representing the conditional probability of treatment given covariates and mediator(s), $\rho(M, X) = \mathbb{P}(T = 1 \mid X, M)$ for the treated and the untreated in representative simulation settings for each of the three (ω_T, ω_M) value sets.



(a) setting 3 (b) setting 15 (c) setting 27

(d) setting 6 (e) setting 18 (f) setting 30

(g) setting 12 (h) setting 24 (i) setting 36

Figure S1: Propensity score for different simulation settings, with low indirect effect size (left column), high indirect effect size and no overlap violation (middle column) or high indirect effect size and strong overlap violation (right column), for a binary one-dimension mediator (first row), a continuous one-dimension mediator (second row) or a continuous five-dimension mediator (last row). As expected, we see the strongest overlap violations for the right column, with a high indirect effect size but driven by a stronger link between T and M (high ω_T coefficient).

S3.1 Theoretical effects

We are interested in the natural direct effect, θ, and in the natural indirect effect, δ. We provide here a computation of the theoretical value of those effects.

S3.1.1 with a binary mediator

$$
\begin{aligned}
\theta(t) &= \mathbb{E}_X\left[\mathbb{E}\left[Y(1,M(t)) - Y(0,M(t))|X=x\right]\right] \\
&= \mathbb{E}_X\Big[\mathbb{E}\Big[\gamma_0 + x^T\gamma_X + \gamma_T + \omega_M\gamma_M M(t) + \gamma_{MT}M(t) \\
&\quad -(\gamma_0 + x^T\gamma_X + \omega_M\gamma_M M(t))|X=x\Big]\Big] \\
&= \mathbb{E}_X\left[\mathbb{E}\left[\gamma_T + \gamma_{MT}M(t)|X=x\right]\right] \\
&= \gamma_T + \gamma_{MT}\mathbb{E}_X\left[\text{expit}(\beta_0 + X^T\beta_X + (tX)^T\beta_{TX} + \omega_T\beta_T t)\right]
\end{aligned}
$$

$$
\begin{aligned}
\delta(t) &= \mathbb{E}_X\left[\mathbb{E}\left[Y(t, M(1)) - Y(t, M(0))|X = x\right]\right] \\
&= \mathbb{E}_X\Big[\mathbb{E}\Big[\gamma_0 + X^T\gamma_X + \gamma_T t + \omega_M\gamma_M M(1) + \gamma_{MT}M(1)t \\
&-(\gamma_0 + X^T\gamma_X + \gamma_T t + \omega_M\gamma_M M(0) + \gamma_{MT}M(0)t)\Big]\Big] \\
&= \mathbb{E}_X\left[\mathbb{E}\left[(M(1) - M(0))(\omega_M\gamma_M + t\gamma_{MT})\right]\right] \\
&= \mathbb{E}_X\Big[\mathbb{E}\Big[(\text{expit}(\beta_0 + X^T\beta_X + X^T\beta_{TX} + \omega_T\beta_T) - \\
&\text{expit}(\beta_0 + X^T\beta_X))(\omega_M\gamma_M + t\gamma_{MT})\Big]\Big]
\end{aligned}
$$

S3.1.2 with a continuous mediator

$$
\begin{aligned}
\theta(t) &= \mathbb{E}_X\left[\mathbb{E}\left[Y(1, M(t)) - Y(0, M(t))|X = x\right]\right] \\
&= \mathbb{E}_X\Big[\mathbb{E}\Big[\gamma_0 + x^T\gamma_X + \gamma_T + \omega_M\gamma_M M(t) + \gamma_{MT}M(t) \\
&-(\gamma_0 + x^T\gamma_X + \omega_M\gamma_M M(t))|X = x\Big]\Big] \\
&= \mathbb{E}_X\left[\mathbb{E}\left[\gamma_T + \gamma_{MT}M(t)|X = x\right]\right] \\
&= \gamma_T + \gamma_{MT}\mathbb{E}_X\left[\beta_0 + X^T\beta_X + (tX)^T\beta_{TX} + \omega_T\beta_T t\right]
\end{aligned}
$$

$$
\begin{aligned}
\delta(t) &= \mathbb{E}_X\left[\mathbb{E}\left[Y(t, M(1)) - Y(t, M(0))|X = x\right]\right] \\
&= \mathbb{E}_X\Big[\mathbb{E}\Big[\gamma_0 + X^T\gamma_X + \gamma_T t + \omega_M\gamma_M M(1) + \gamma_{MT}M(1)t \\
&-(\gamma_0 + X^T\gamma_X + \gamma_T t + \omega_M\gamma_M M(0) + \gamma_{MT}M(0)t)\Big]\Big] \\
&= \mathbb{E}_X\left[\mathbb{E}\left[(M(1) - M(0))(\omega_M\gamma_M + t\gamma_{MT})\right]\right] \\
&= \mathbb{E}_X\Big[\mathbb{E}\Big[\Big((\beta_0 + X^T\beta_X + (X)^T\beta_{TX} + \omega_T\beta_T) - \\
&(\beta_0 + X^T\beta_X)\Big)(\omega_M\gamma_M + t\gamma_{MT})\Big]\Big]
\end{aligned}
$$

S4 How many bootstrap repetitions are required to obtain decent percentile confidence intervals?

To assess a reasonable number of bootstrap repetitions to assess the characteristics of estimators' convergence, we have conducted 1,000 bootstrap repetitions on 100 varied simulated datasets. We represent the mean, and the 2.5 and 97.5 percentiles for different number of bootstrap samples ranging from 1 to 1,000 in Supplementary Figure S2.

S5 Extended results on simulations

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

We consider the variant estimators listed in Table S2.

Supplementary Figure S4 represents the coverage, and Supplementary Figure S5 the width of 95% percentile bootstrap confidence intervals for the estimation of the different effects. We observe results that are very similar to what is obtained for the estimation error. A valuable insight is that the IPW estimator systematically has wider confidence intervals compared to the other methods, and that the doubly-robust estimators has slightly narrower intervals than the G-computation approach. This can constitute a discriminative criterion for the choice of the estimator as both estimators have very similar results for the estimation error.

In Supplementary Figure S3, we present the same results as Figure 3 with the continuous one-dimensional mediator (second row) and the multi-dimensional mediator (third row). The pattern of bias is similar to the pattern observed for the binary mediator, with an important bias of the IPW estimator when the mediator model is misspecified, and for the coefficient product when the outcome model is misspecified. In the case

Coefficient product
G-computation (forest)
G-computation (forest & cross-fitting)
G-computation (no regularization)
G-computation (regularization)
G-computation (regularization & cross-fitting)
G-computation (forest)
G-computation (forest & cross-fitting)
G-computation (no regularization)
G-computation (regularization)
G-computation (regularization & cross-fitting)
IPW (forest)
IPW (forest & cross-fitting)
IPW (no regularization)
IPW (regularization)
IPW (regularization & cross-fitting)
multiply robust (forest)
multiply robust (forest & cross-fitting)
multiply robust (no regularization)
multiply robust (regularization)
multiply robust (regularization & cross-fitting)
medDML (no regularization)
medDML (regularization)
medDML (regularization & cross-fitting)
medDML (forest)
medDML (forest & cross-fitting)

Table S2: **List of considered estimators and their implementation variants**

of a multi-dimensional mediator, the bias is less important for the IPW estimator when the mediator model is misspecified. The higher number of mediators may provide additional intrinsic flexibility to the propensity score model, as the mediators are used solely to predict the binary treatment and are not explicitly modeled themselves. The G-computation estimator does not exhibit a bias in the estimation when the mediator is continuous. This is related to our implementation for the case of the implicit integration for the mediator density involves two distinct regressions for the outcome model, one regression for the treated units, and one for the untreated units. This pattern is similar to the difference between the S-learner and the T-learner estimators in standard causal inference without mediation, and is known to be able to model properly interactions between the treatment and the covariates [Künzel et al., 2019]. This property explains why the G-computation in this implementation does not suffer from the outcome model misspecification. However, in theory, it would handle badly more complex non-linearities of the covariates.

In Supplementary Figure S6, we present the same results as Figure 4 with the absolute error, so that the reader can fully interpret the simulation errors, and include results for the continuous mediators. The IPW estimators exhibits an important error rate for continuous mediators and overlap violation even in the absence of misspecification of the mediator model.

We also consider the behavior of the confidence interval coverage and width over those different configurations in Supplementary Figures S7 and S8. In this case, the behavior of those additional performance metrics is very different from the estimation error. Indeed, the double machine learning estimator with multi-dimensional mediators exhibits a very important drop in coverage when there is an overlap violation, but only a modest estimation error, that does not appear statistically significant. For the estimators involving inverse-propensity-weighting (IPW, G-computation and) continuous mediators, we observe an important increase

In the following experiments, we focus on four estimators that we have re-implemented in Python for this study, and assess the impact of using machine learning for the nuisance functions, instead of plain parametric linear models. We consider the IPW estimator, the G-computation estimator, the multiply-robust estimator and the double-machine learning estimator. A first improvement we can apply is regularization, by imposing a penalty on the size of the coefficients. In our implementation, the hyperparameter setting the amount of shrinkage is set by cross-validation. Supplementary Figure S9 presents the results for the unregularized and the regularized versions of those three estimators in the "high indirect effect size without overlap violation" framework. Overall, the regularized estimators underperform their unregularized equivalent. Contrary to what we observed with the unregularized estimators (Figure 5, and Supplementary Figure S9), we see that the estimation error decreases with the number of samples, indicating a slow convergence. However, even with a large sample, the error is still

much larger for the regularized estimators. This could be explained by the fact that this is a well-posed regime, with not-so-high dimensional data and a lot of samples. Regularization becomes futile and may bias estimators when we compare different parameter settings.

We observe the same trend, and no improvement on the estimation accuracy when we replace the nuisance functions models by random forests [Breiman, 2001] (Supplementary Figure S10, even when the mediator or outcome models are misspecified. In the case of the G-computation, with a very large number of observations, the performance of the estimator with forest nuisance models seems to catch up with the linear model, indicating that it could eventually converge to an unbiased estimation for the one-dimensional binary or continuous mediator, however, the convergence is really slow for the multi-dimensional case. For the double machine learning estimator, the robustness property with misspecified nuisance function seems to be the best solution.

In Supplementary Figures S11 and S12, we explore the potential improvement obtained by using cross-fitting for the nuisance models. This is generally presented as a solution to the slow convergence of machine learning models for the nuisance parameters Chernozhukov et al. [2018]. However, we observe rather a slightly slower convergence, probably due to the smaller sample size.

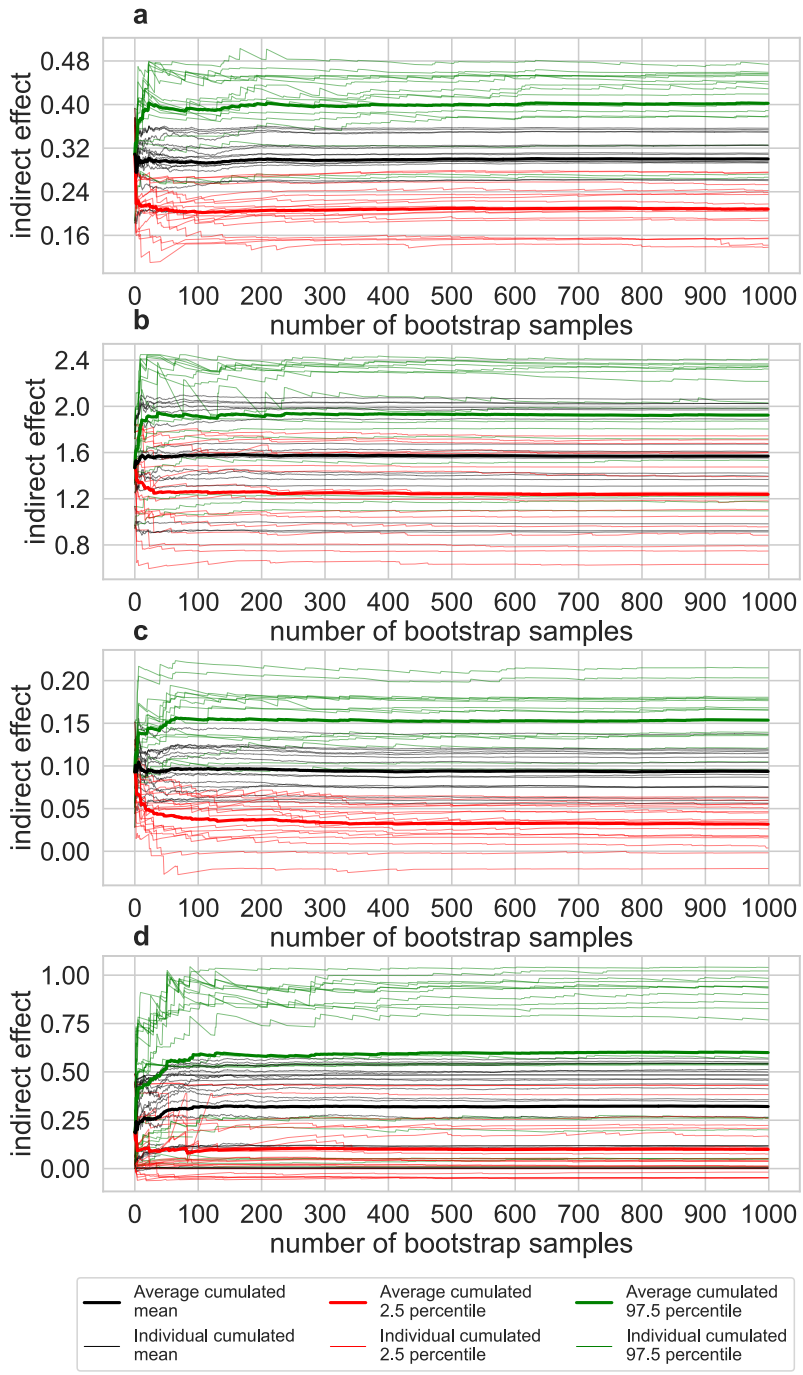


Figure S2: **Exploration of an adequate number of bootstrap samples for uncertainty quantification** We show the value of the bounds of the 95%-percentile confidence interval for different numbers of bootstrap samples. The bold curves represent the average over the 26 individual estimators, and the thin lines are the results for each estimator. Each panel represents a different simulation configuration: **a** 500 observations, with a one-dimensional continuous mediator, in the setting of a low indirect effect size without overlap violation, and a misspecification of the outcome model; **b** 500 observations, with a one-dimensional continuous mediator, in the setting of a high indirect effect size with overlap violation, and a misspecification of both the outcome and the mediator models; **c** 1,000 observations, with a five-dimensional continuous mediator, in the setting of a low indirect effect size without overlap violation, and a misspecification of the outcome model; **d** 1,000 observations, with a binary mediator, in the setting of a high indirect effect size with overlap violation, and a misspecification of the mediator model. This experiment shows that most estimators have reached a stable point with 100 bootstrap samples.

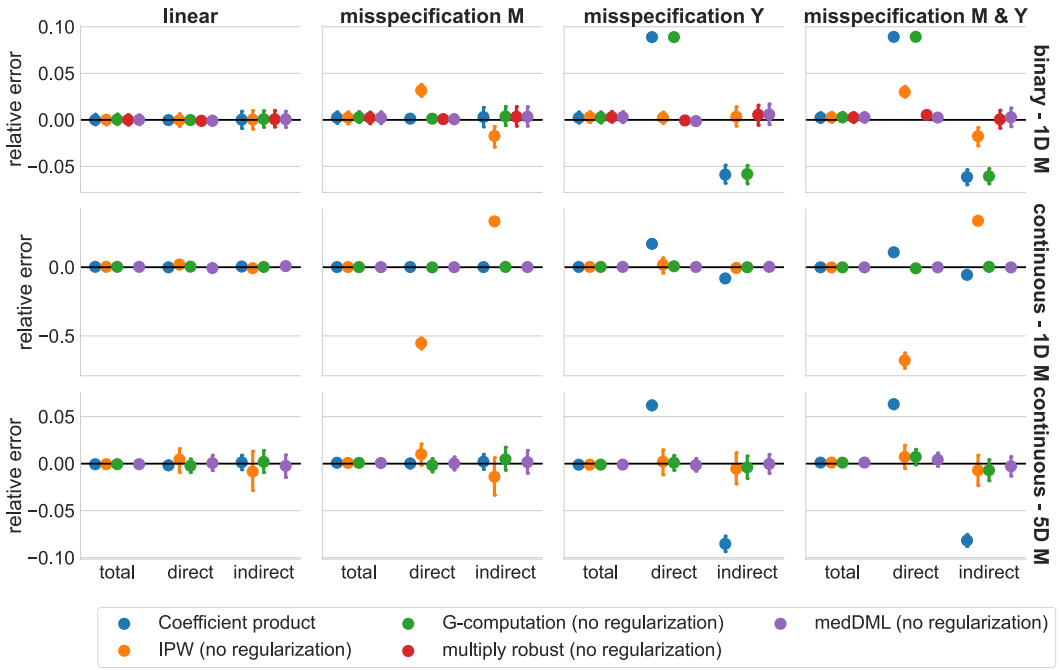


Figure S3: **Total and natural direct and indirect effects obtained with unregularized estimators for all mediator types.** We present the results of four generative model specification scenarios, some of which violate parametric linear nuisance models for certain estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, 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 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the *high indirect effect size without overlap violation* framework with $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows. The total effect is generally well estimated in all situations, but not the direct and indirect effects. The indirect and direct effects have different value, leading to a lack of symmetry in their relative error values, even if their absolute values are equal, leading to no error for the total effect estimation. Model misspecifications lead to estimation errors for most estimators, especially when the outcome model is misspecified.

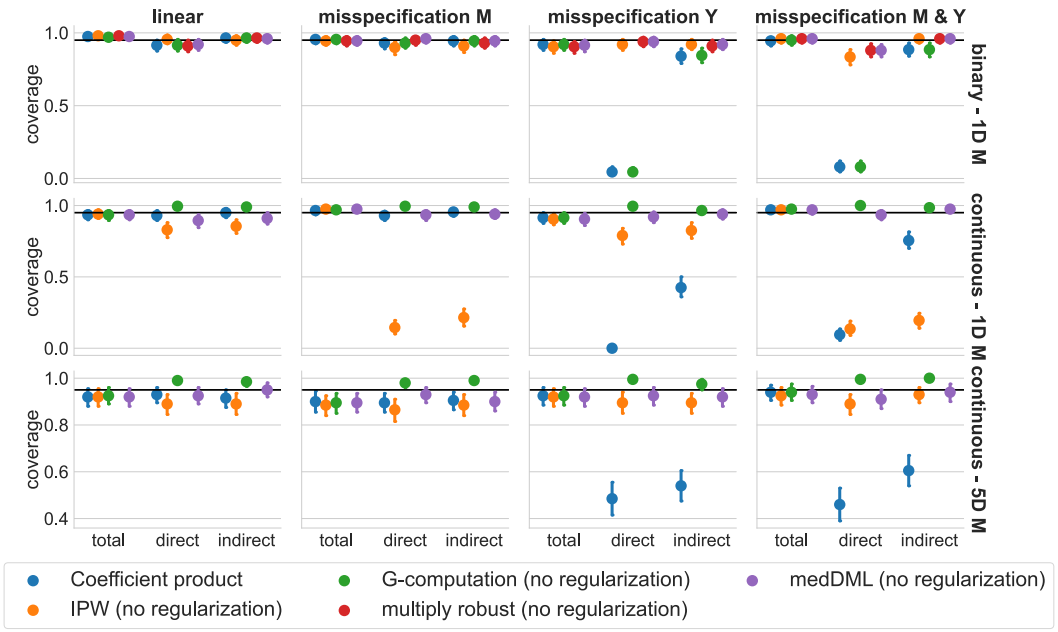


Figure S4: **95%-confidence-interval coverage for the total and natural direct and indirect effects obtained with unregularized estimators for all mediator types.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. Each dot represents the average coverage obtained with 100 bootstrap samples, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the *high indirect effect size without overlap violation* framework with $n = 1,000$ observations.

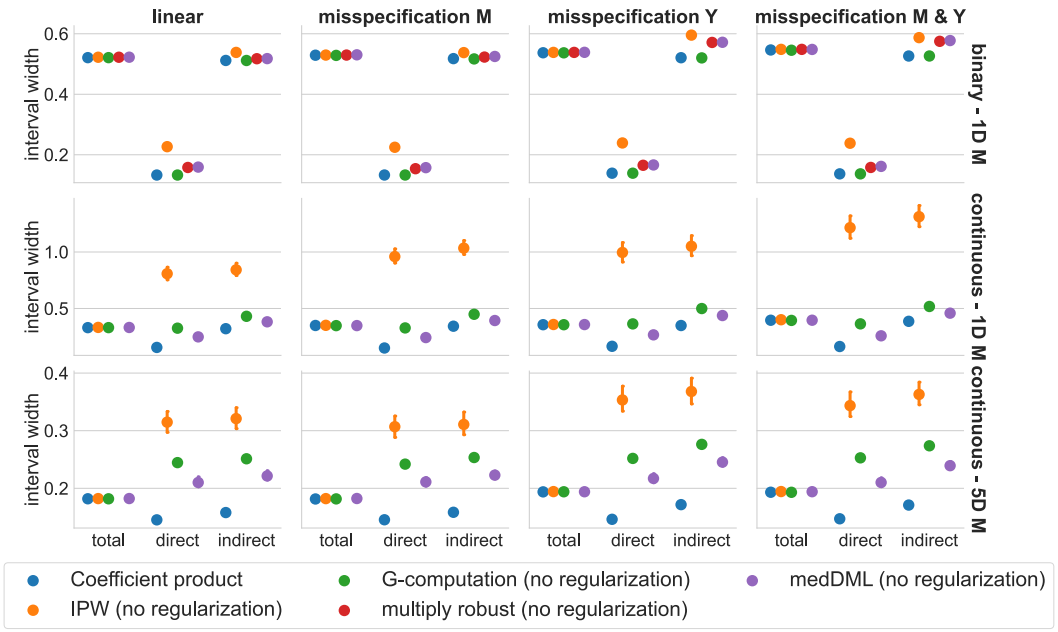


Figure S5: **95%-confidence-interval width for the total and natural direct and indirect effects obtained with unregularized estimators for all mediator types.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. Each dot represents the average interval width obtained with 100 bootstrap samples, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the *high indirect effect size without overlap violation* framework with $n = 1,000$ observations.

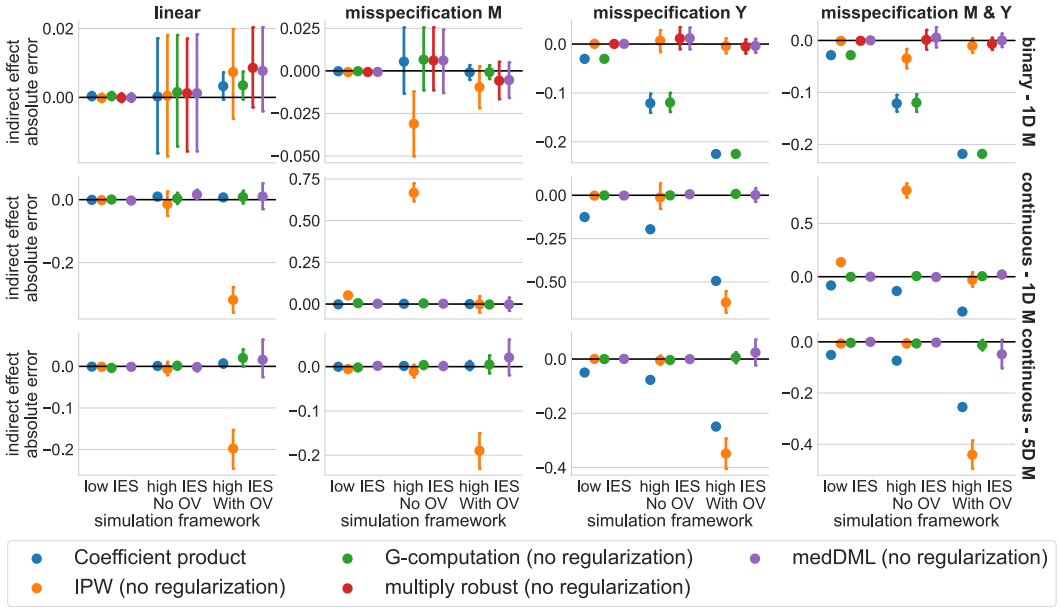


Figure S6: **Indirect absolute effect for different simulation frameworks.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average relative error (i.e. $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows. The estimation error for most estimators is greater when there is a violation of the overlap assumption, including in robust estimators. The absolute value of the indirect effect in the "low indirect effect size without overlap violation" is often very low, leading to very high relative errors compared to other settings. We found however that the representation of the relative error in Figure 4 is informative. We represent here the same results as Figure 4 with the absolute error (and not relative).

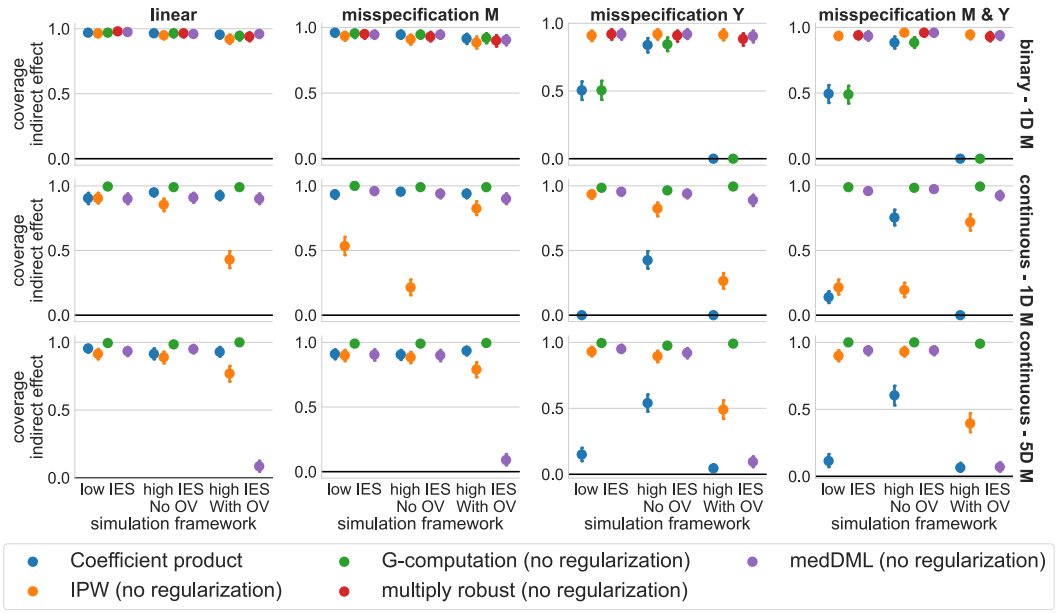


Figure S7: **Coverage for the indirect effect for different simulation frameworks.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average coverage over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows.

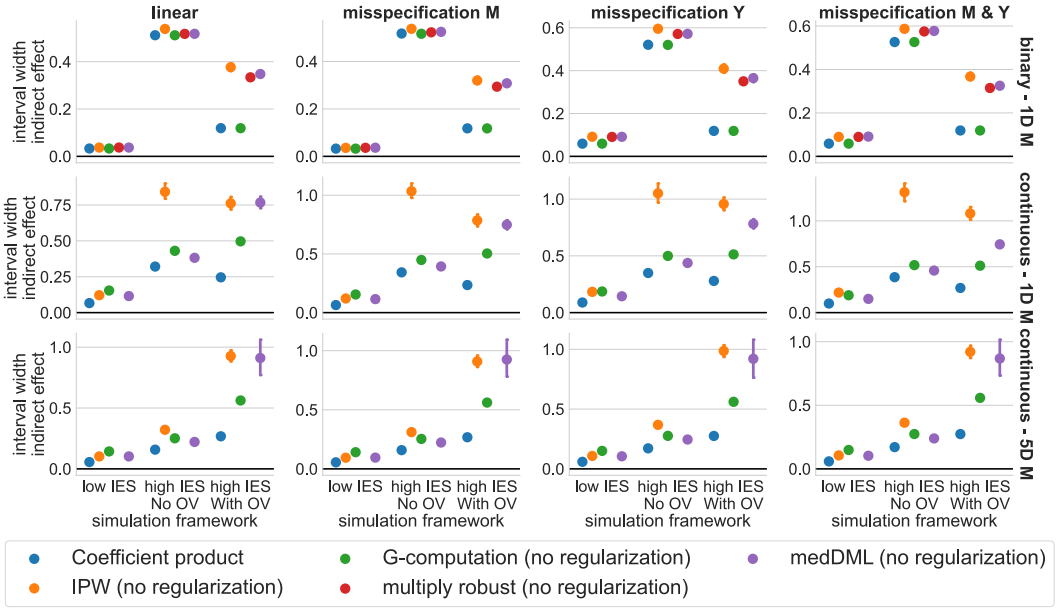


Figure S8: **Confidence interval width for the indirect absolute effect for different simulation frameworks.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average interval width over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows.

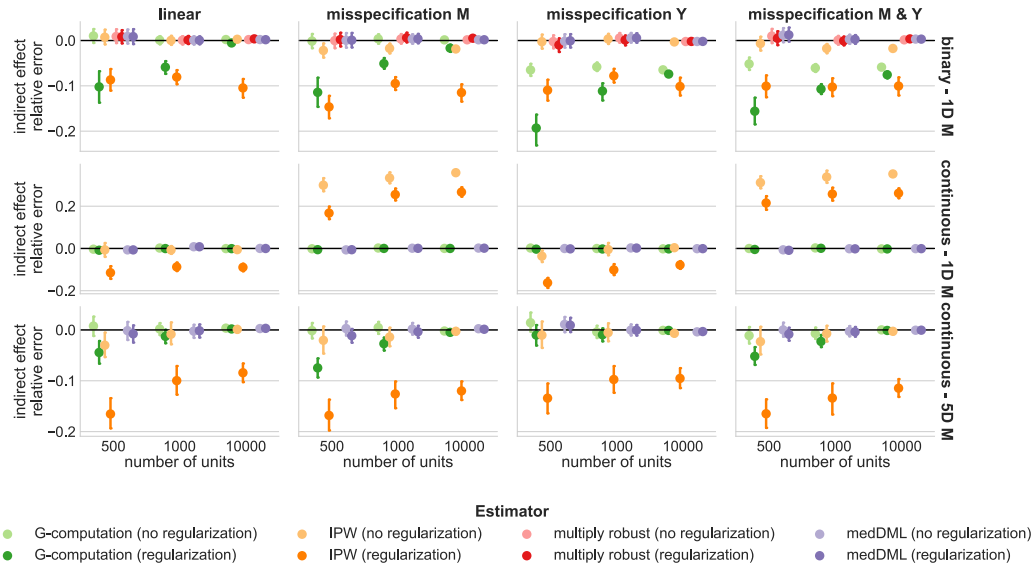


Figure S9: **Indirect effect relative error with regularized and unregularized linear models for plug-in nuisance parameters estimation.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The four sets of dots in each panel correspond to the different number of observations tested in our set of simulations. Each dot represents the average relative error (i.e. $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the "high indirect effect size without overlap violation" framework. We focus on four estimators that were re-implemented in Python for this study, and assess the impact of using regularized models for the nuisance functions. In all simulated cases, it appears preferable to use unregularized models.

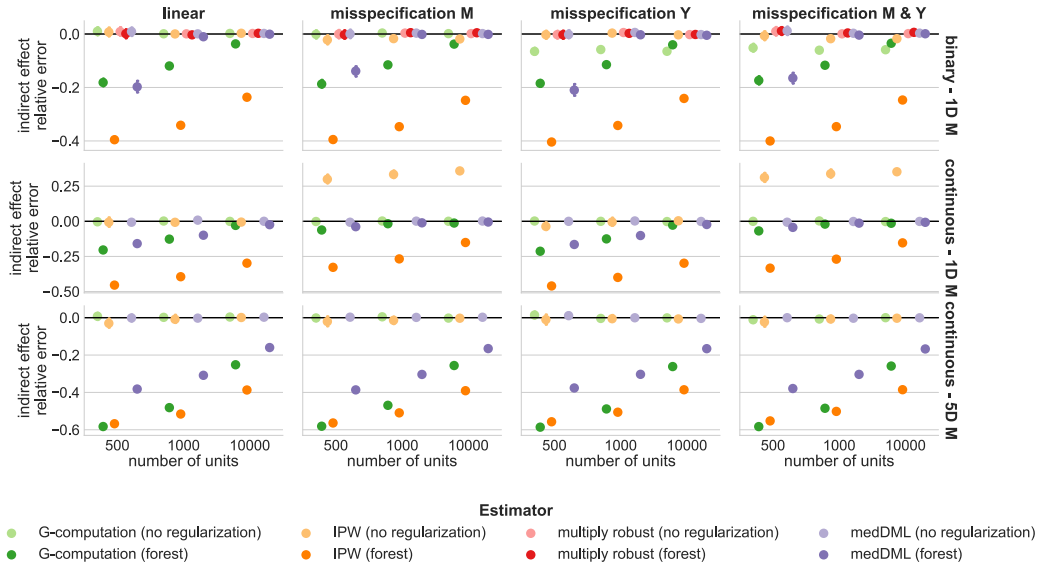


Figure S10: **Indirect effect relative error with random forests for plug-in nuisance parameters estimation.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The three sets of dots in each panel correspond to the different number of observations tested in our set of simulations. Each dot represents the average relative error (i.e. $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the "high indirect effect size without overlap violation" framework.

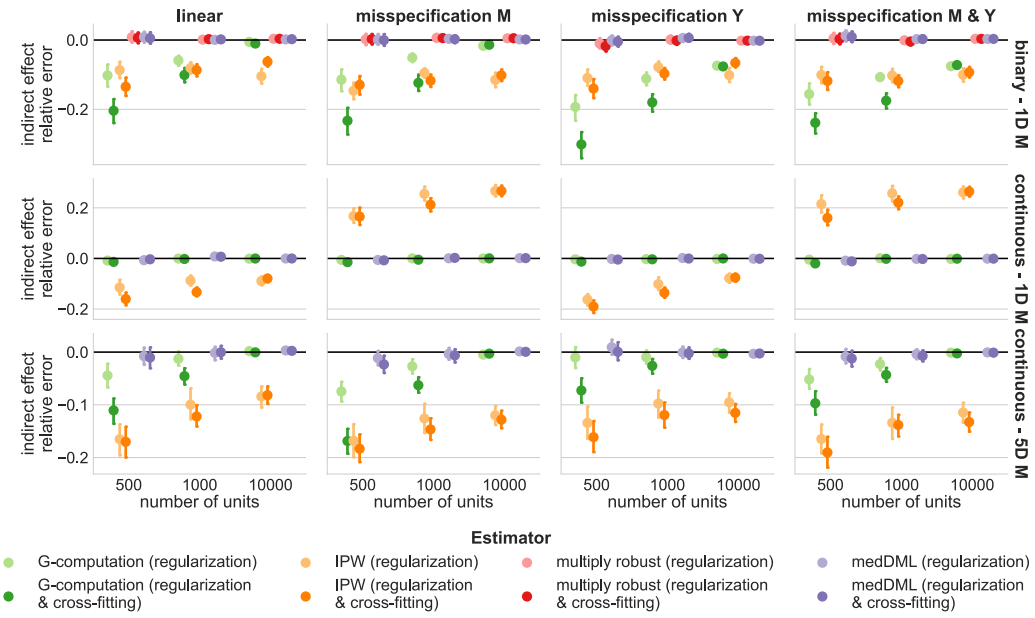


Figure S11: **Indirect effect relative error with linear models with and without cross-fitting for plug-in nuisance parameters estimation.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The four sets of dots in each panel correspond to the different number of observations tested in our set of simulations. Each dot represents the average relative error (i.e. $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the "high indirect effect size without overlap violation" framework.

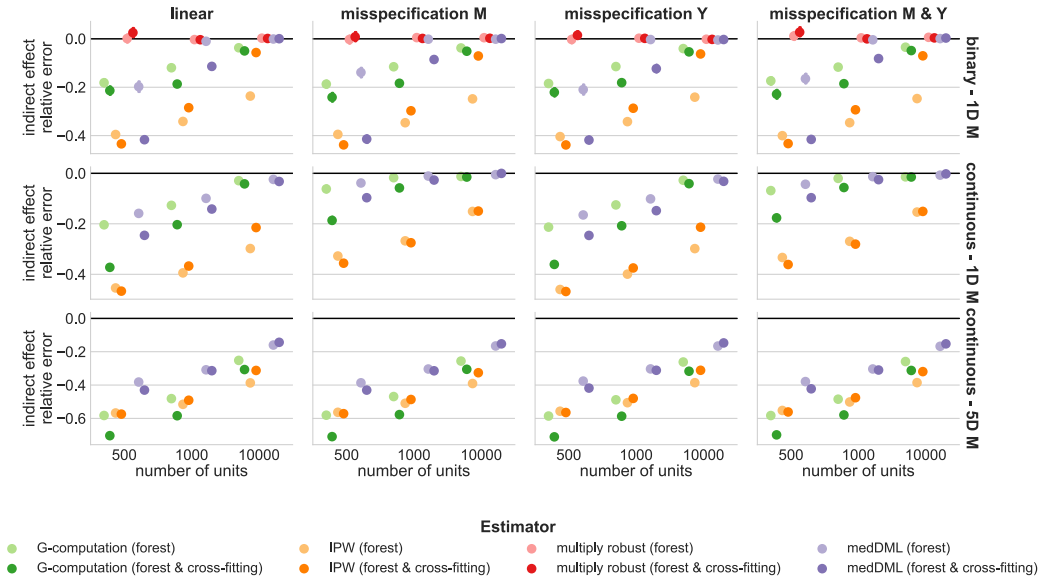


Figure S12: **Indirect effect relative error with random forests models with and without cross-fitting for plug-in nuisance parameters estimation.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models of some estimators. Each column corresponds to a distinct specification of simulated models. The rows correspond to different mediator types, labeled on the right. The four sets of dots in each panel correspond to the different number of observations tested in our set of simulations. Each dot represents the average relative error (i.e. $\frac{\hat{\tau}-\tau}{\tau}$, for example for the total effect τ) over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the "high indirect effect size without overlap violation" framework.

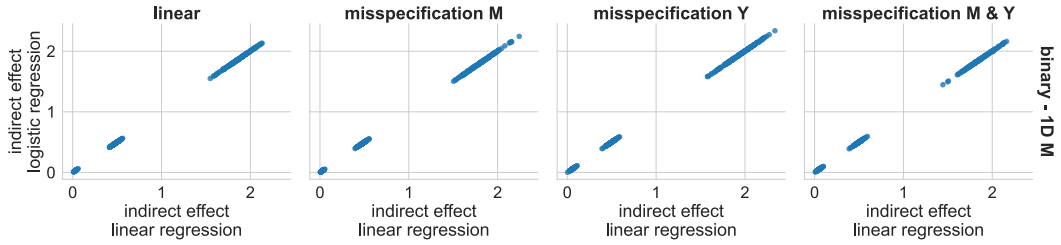


Figure S13: **Indirect effect estimation for two alternative implementations coefficient product: linear regression or logistic regression for the mediator model in the binary case.** We show results for four scenarios of generative model specification, violating or not the parametric linear nuisance models for the case of a binary mediator for the coefficient product. Each column corresponds to a distinct specification of simulated models. The two axes represent the estimates obtained using a linear regression model for the binary mediator (x-axis, this is the setting represented in all the other experiments of this work), and using a logistic regression, which in theory is better suited for a binary mediator. Each dot represents a simulated dataset. All the configurations with different indirect effect sizes and overlap violations or not, and sample sizes are represented together. In all situations, the estimates obtained with those two approaches are exactly the same, validating the simplification that was made to avoid further increasing the number of estimators considered.

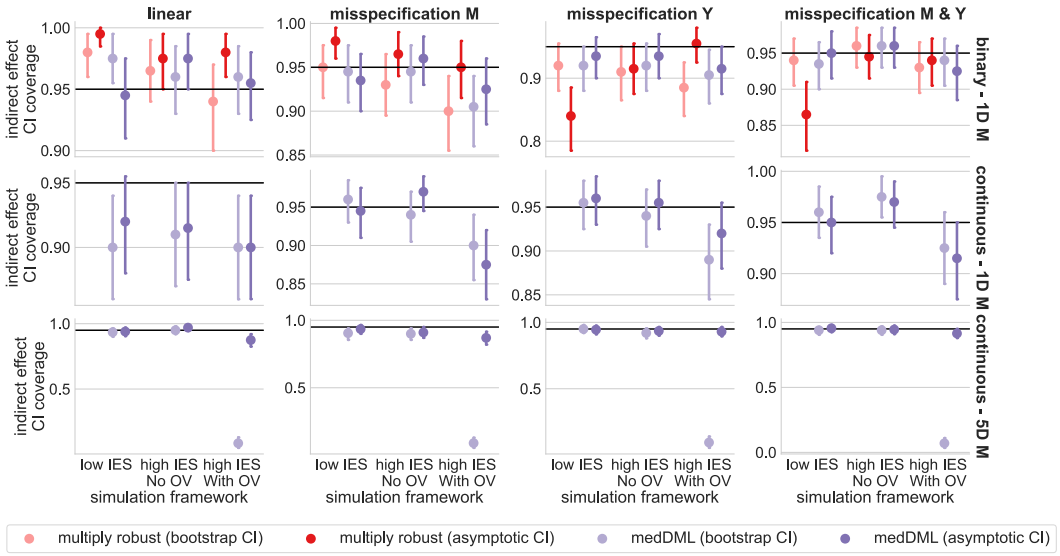


Figure S14: **Coverage for the indirect effect for different simulation frameworks, for the bootstrap and the asymptotic confidence intervals for the multiply robust and double machine learning estimators, with linear models for the nuisance functions.** We show results for four scenarios of generative 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 types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average coverage over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows. We compare two strategies to obtain confidence intervals. Both bootstrap and asymptotic confidence intervals exhibit a good coverage in all situations, except for the bootstrap confidence intervals in the presence of multi-dimensional mediators with overlap violation.

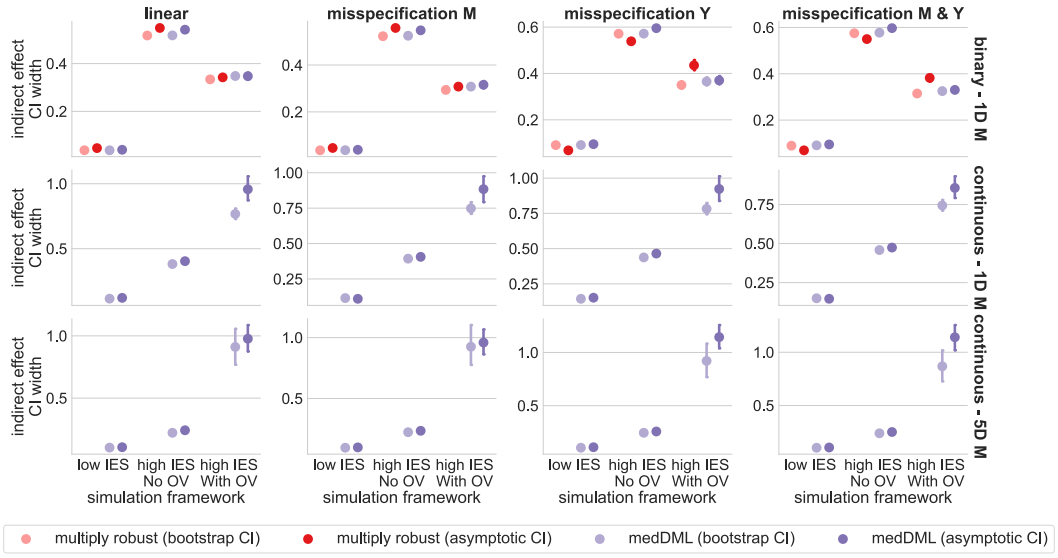


Figure S15: **Confidence interval width for the indirect effect for different simulation frameworks, for the bootstrap and the asymptotic confidence intervals for the multiply robust and double machine learning estimators, with linear models for the nuisance functions.** We show results for four scenarios 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 types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average coverage over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows. We compare two strategies to obtain confidence intervals. Both bootstrap and asymptotic confidence intervals exhibit similar widths in all situations, except for the bootstrap confidence intervals in the presence of multi-dimensional mediators with overlap violation.

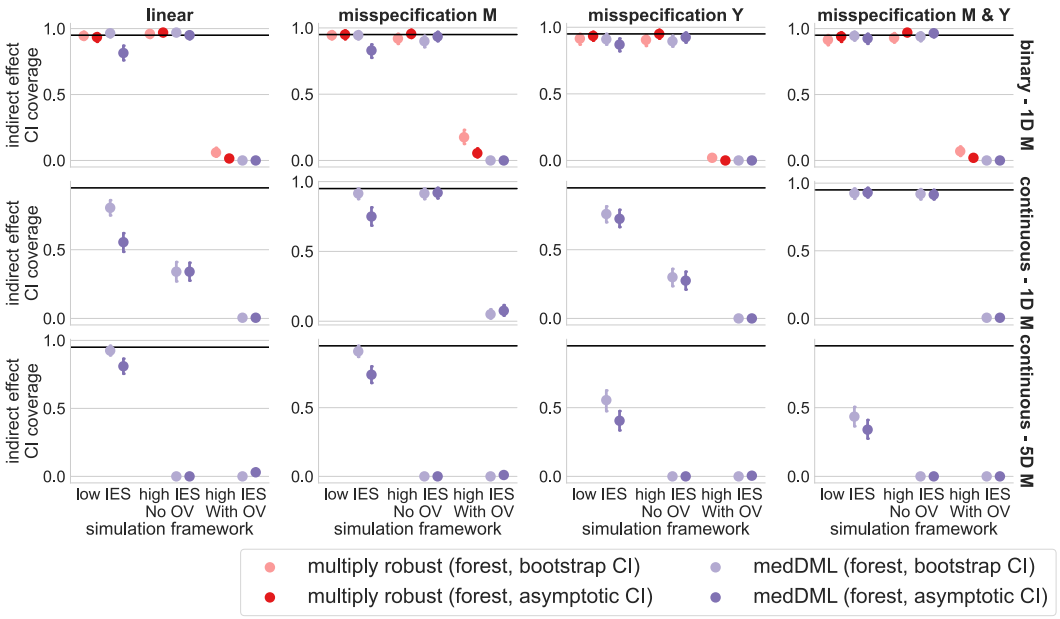


Figure S16: **Coverage for the indirect effect for different simulation frameworks, for the bootstrap and the asymptotic confidence intervals for the multiply robust and double machine learning estimators, with random forests models for the nuisance functions.** We show results for four scenarios of 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 types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average coverage over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows. We compare two strategies to obtain confidence intervals. Both bootstrap and asymptotic confidence intervals achieve a similar coverage in all situations. This coverage is however systematically very bad when there is a violation of the overlap assumption, where the estimator is not supposed to be valid. We also see bad coverage performance with continuous or multi-dimensional mediators, which is a case where the estimator has actually not converged based on the results shown in Supplementary Figure S10.

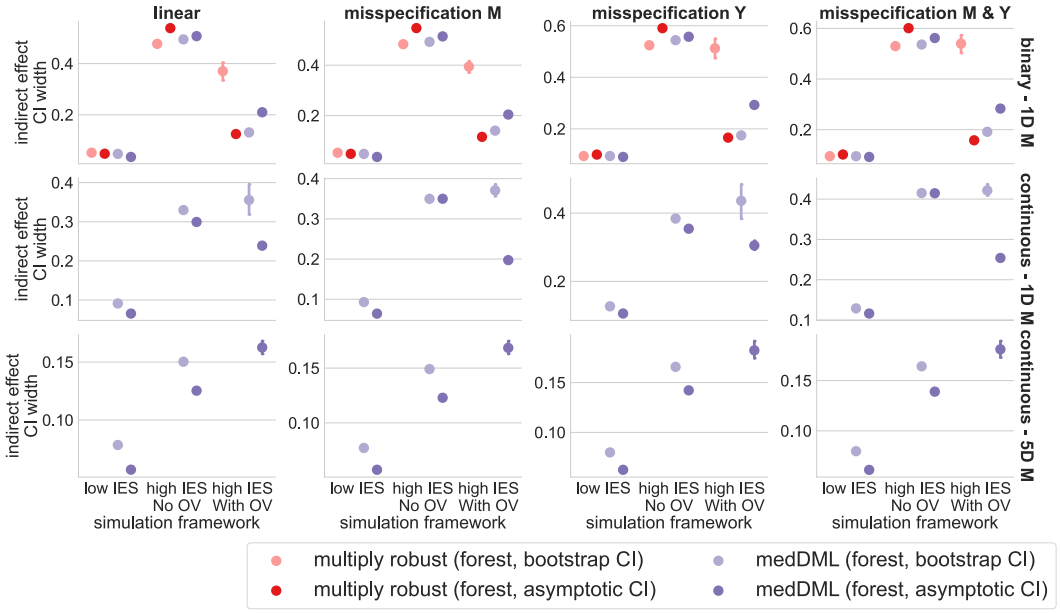


Figure S17: **Confidence interval width for the indirect effect for different simulation frameworks, for the bootstrap and the asymptotic confidence intervals for the multiply robust and double machine learning estimators, with random forests models for the nuisance functions.** We show results for four scenarios 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 types, labeled on the right. The three sets of dots in each panel correspond to the three simulation frameworks, in that order: "low indirect effect size without overlap violation", "high indirect effect size without overlap violation" and "high indirect effect size with strong overlap violation". "IES" stands for "indirect effect size" and "OV" for "overlap violation". Each dot represents the average coverage over 200 repetitions, and the error bars are the asymptotic normal 95% confidence intervals from the distribution over the 200 repetitions. All simulations are in the case where $n = 1,000$ observations. The multiply-robust estimator only handles binary one-dimensional mediators, and has no results for the second and third rows. We compare two strategies to obtain confidence intervals. Both bootstrap and asymptotic confidence intervals exhibit close behaviors in all situations. In the case of multi-dimensional mediators, the asymptotic confidence interval is slightly narrower, but achieves bad coverage. The bootstrap confidence intervals have failed for the case of multi-dimensional mediators with overlap violation.

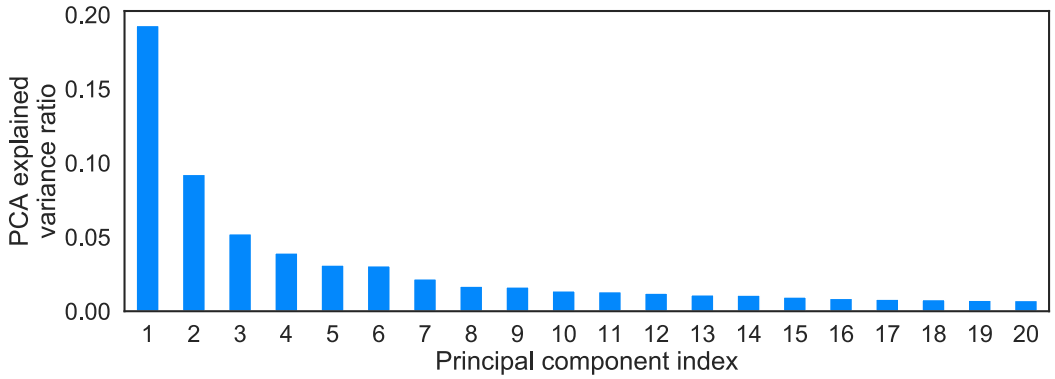


Figure S18: **Explained variance of PCA components of sMRI and dMRI IDPs (Image Derived Phenotypes).** This barplot shows the percentage of the dataset variation that is explained by each component of a PCA applied to the 311 retained IDPs variables from the raw brain MRI data. We retained 6 components after visual inspection.

Table S3: **Description of aggregated variables.** This table describes how several fields from the raw UKBB data are combined together to define binary variables, as treatment or confounder.

exposure	UBKK fields	aggregation rationale
college or university degree	**6138** (Qualifications)	equal to 'College or University degree'
diabetic	fields **20002** (Non-cancer illness code, self-reported) and **6177** (Medication for cholesterol, blood pressure or diabetes) and **6153** (Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones) at the imaging visit	self-declared diabetes or medication for diabetes, inspired from Newby and Garfield [2022]
hypercholesterol	fields **20002** (Non-cancer illness code, self-reported) and **6177** (Medication for cholesterol, blood pressure or diabetes) and **6153** (Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones) at the imaging visit	self-declared hypercholesterol or medication for hypercholesterol, inspired from Newby and Garfield [2022]
hypertension	fields **20002** (Non-cancer illness code, self-reported) and **6177** (Medication for cholesterol, blood pressure or diabetes) and **6153** (Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones) or `6150` (Vascular/heart problems diagnosed by doctor) at the imaging visit	self-declared hypertension or medication for hypertension, or diagnosed by a medical doctor, inspired from Newby and Garfield [2022]
Current MDD (Major depressive disorder)	fields **2050** (Frequency of depressed mood in last 2 weeks), **2060** (Frequency of unenthusiasm disinterest in last 2 weeks), **4526** (Happiness), **1930** (Miserableness), **1960** (Fed-up feelings), **1170** (Getting up in morning), **1200** (Sleeplessness insomnia), **2070** (Frequency of tenseness restlessness in last 2 weeks), **2080** (Frequency of tiredness lethargy in last 2 weeks), **1940** (Irritability), **1950** (Sensitivity hurt feelings), **2020** (Loneliness, isolation), **2030** (Guilty feelings) at the imaging visit	combination criterion inspired from Stolicyn et al. [2020]
Lifetime MDD	fields **20441** (Ever had prolonged loss of interest in normal activities), **20446** (Ever had prolonged feelings of sadness or depression), **20532** (Did your sleep change?), **20435** (Difficulty concentrating during worst depression), **20449** (Feelings of tiredness during worst episode of depression), **20450** (Feelings of worthlessness during worst period of depression), **20437** (Thoughts of death during worst depression), **20536** (Weight change during worst episode of depression), **20436** (Fraction of day affected during worst episode of depression), **20439** (Frequency of depressed days during worst episode of depression), **20440** (Impact on normal roles during worst period of depression) at the initial assessment	combination criterion inspired from Stolicyn et al. [2020]
social support	field **2110** (Able to confide) at the imaging visit	threshold similar to Schurz et al. [2021]
inflammation	field **30710** (C-reactive protein) at the initial visit	Based on the expected range, we took a cutoff of 3 mg/L as a marker of potential chronic inflammation [Pepys et al., 2003, Pearson et al., 2003].
current smoking	fields **1239** (Current tobacco smoking) and **1249** (Past tobacco smoking) at imaging visit	self declaration
alcohol consumption	fields **1568** (Average weekly red wine intake), **1578** (Average weekly champagne plus white wine intake), **1588** (Average weekly beer plus cider intake), **1598** (Average weekly spirits intake), **1608** (Average weekly fortified wine intake), **5364** (Average weekly intake of other alcoholic drinks) at imaging visit	the different alcoholic drinks were converted to standard units as done in the work of Topiwala et al. [2022]. We set a cutoff at the recommended level by the National Health Service (NHS) of less than 14 units per week.
obesity	**21001** (Body mass index BMI) at imaging visit	the standard cutoff of the BMI, that is greater than 30 kg/m^2 was used.

Table S4: **Baseline characteristics of for the included participants, according to their higher hypertensive status**. Adequate statistical tests depending on the variable type assess whether each variable has a distinct distributions between the "exposed" and the "unexposed" group (with exposure having hypertension). We also report the Standardized mean difference (SMD) for each variable.

variable role	variable name	variable modalities	Overall	No hypertension	Hypertension	P-Value	Test	SMD (0,1)
Sample size	n		32498	23962	8536			
Outcome	G factor, mean (SD)		-0.0 (1.1)	-0.1 (1.1)	0.1 (1.2)	<0.001	Two Sample T-test	0.169
Mediators	Dmri pc 0, mean (SD)		-0.3 (10.2)	-1.6 (9.3)	3.4 (11.5)	<0.001	Two Sample T-test	0.476
	Dmri pc 1, mean (SD)		0.1 (5.2)	0.4 (5.1)	-0.8 (5.5)	<0.001	Two Sample T-test	-0.225
	Dmri pc 2, mean (SD)		-0.1 (5.0)	-0.4 (4.9)	0.8 (5.2)	<0.001	Two Sample T-test	0.241
	Dmri pc 3, mean (SD)		0.0 (4.4)	-0.3 (4.3)	0.7 (4.5)	<0.001	Two Sample T-test	0.226
	Dmri pc 4, mean (SD)		0.0 (4.2)	0.2 (4.1)	-0.4 (4.7)	<0.001	Two Sample T-test	-0.140
	Dmri pc 5, mean (SD)		0.0 (3.4)	0.1 (3.3)	-0.2 (3.7)	<0.001	Two Sample T-test	-0.061
	Smri dmri pc 0, mean (SD)		-0.4 (10.5)	-1.8 (9.6)	3.6 (11.8)	<0.001	Two Sample T-test	0.497
	Smri dmri pc 1, mean (SD)		-0.2 (7.4)	-0.1 (7.4)	-0.4 (7.5)	0.006	Two Sample T-test	-0.035
	Smri dmri pc 2, mean (SD)		0.1 (5.6)	0.5 (5.4)	-1.1 (5.9)	<0.001	Two Sample T-test	-0.284
	Smri dmri pc 3, mean (SD)		-0.0 (4.7)	-0.2 (4.6)	0.5 (4.9)	<0.001	Two Sample T-test	0.150
	Smri dmri pc 4, mean (SD)		-0.0 (4.3)	-0.3 (4.3)	0.7 (4.4)	<0.001	Two Sample T-test	0.223
	Smri dmri pc 5, mean (SD)		0.0 (4.3)	0.2 (4.1)	-0.5 (4.7)	<0.001	Two Sample T-test	-0.170
	Smri pc 0, mean (SD)		0.3 (7.2)	0.6 (7.1)	-0.7 (7.2)	<0.001	Two Sample T-test	-0.175
	Smri pc 1, mean (SD)		-0.0 (3.3)	-0.3 (3.2)	0.8 (3.5)	<0.001	Two Sample T-test	0.326
	Smri pc 2, mean (SD)		-0.1 (2.8)	-0.4 (2.7)	0.8 (3.0)	<0.001	Two Sample T-test	0.424
	Smri pc 3, mean (SD)		0.0 (2.1)	-0.1 (2.0)	0.2 (2.1)	<0.001	Two Sample T-test	0.146
Confounders (alphabetic order)	Adopted as a child - visit 0, n (%)	No	32469 (99.9)	23939 (99.9)	8530 (99.9)	0.786	Chi-squared (warning: expected count < 5)	0.009
		Yes	9 (0.0)	7 (0.0)	2 (0.0)			
		other	20 (0.1)	16 (0.1)	4 (0.0)			
	Age when attended assessment centre - visit 2, mean (SD)		63.7 (7.6)	62.7 (7.5)	66.4 (7.1)	<0.001	Two Sample T-test	0.502
	Alcohol above nhs reco, n (%)	0	21256 (65.4)	15968 (66.6)	5288 (61.9)	<0.001	Chi-squared	0.098
		1	11242 (34.6)	7994 (33.4)	3248 (38.1)			
	Alcohol intake frequency. - visit 2, n (%)	Daily or almost daily	5540 (17.0)	3890 (16.2)	1650 (19.3)	<0.001	Chi-squared	0.116
		Never	2032 (6.3)	1464 (6.1)	568 (6.7)			
		Once or twice a week	8640 (26.6)	6611 (27.6)	2029 (23.8)			
		One to three times a month	3788 (11.7)	2818 (11.8)	970 (11.4)			
		Special occasions only	3262 (10.0)	2323 (9.7)	939 (11.0)			
		Three or four times a week	9236 (28.4)	6856 (28.6)	2380 (27.9)			
	Average total household income before tax - visit 2, n (%)	18,000 to 30,999	7813 (24.0)	5488 (22.9)	2325 (27.2)	<0.001	Chi-squared	0.188
		31,000 to 51,999	9156 (28.2)	6707 (28.0)	2449 (28.7)			
		52,000 to 100,000	7028 (21.6)	5495 (22.9)	1533 (18.0)			
		Do not know	961 (3.0)	693 (2.9)	268 (3.1)			
		Greater than 100,000	2277 (7.0)	1847 (7.7)	430 (5.0)			
		Less than 18,000	3402 (10.5)	2383 (9.9)	1019 (11.9)			
		Prefer not to answer	1861 (5.7)	1349 (5.6)	512 (6.0)			
	Body mass index (bmi) - visit 2, mean (SD)		26.4 (4.3)	25.8 (4.1)	28.0 (4.6)	<0.001	Two Sample T-test	0.505
	Breastfed as a baby - visit 0, n (%)	Do not know	6175 (19.0)	4231 (17.7)	1944 (22.8)	<0.001	Chi-squared (warning: expected count < 5)	0.159
		No	7174 (22.1)	5606 (23.4)	1568 (18.4)			
		Yes	19136 (58.9)	14115 (58.9)	5021 (58.8)			
		other	13 (0.0)	10 (0.0)	3 (0.0)			
	Cmdd final, n (%)	False	23413 (72.0)	17295 (72.2)	6118 (71.7)	0.381	Chi-squared	0.011
		True	9085 (28.0)	6667 (27.8)	2418 (28.3)			
	Comparative body size at age 10 - visit 0, n (%)	About average	16637 (51.2)	12424 (51.8)	4213 (49.4)	<0.001	Chi-squared (warning: expected count < 5)	0.063
		Do not know	347 (1.1)	239 (1.0)	108 (1.3)			
		Plumper	4922 (15.1)	3518 (14.7)	1404 (16.4)			
		Thinner	10584 (32.6)	7775 (32.4)	2809 (32.9)			
		other	8 (0.0)	6 (0.0)	2 (0.0)			

...continued

variable role	variable name	variable modalities	Overall	No hypertension	Hypertension	P-Value	Test	SMD (0,1)
	Comparative height size at age 10 - visit 0, n (%)	About average	17039 (52.4)	12478 (52.1)	4561 (53.4)	0.032	Chi-squared (warning: expected count < 5)	0.041
		Do not know	293 (0.9)	201 (0.8)	92 (1.1)			
		Shorter	6569 (20.2)	4861 (20.3)	1708 (20.0)			
		Taller	8589 (26.4)	6416 (26.8)	2173 (25.5)			
		other	8 (0.0)	6 (0.0)	2 (0.0)			
	Country of birth (uk/elsewhere) - visit 0, n (%)	Elsewhere	1778 (5.5)	1331 (5.6)	447 (5.2)	0.559	Chi-squared (warning: expected count < 5)	0.028
		England	27360 (84.2)	20162 (84.1)	7198 (84.3)			
		Northern Ireland	234 (0.7)	179 (0.7)	55 (0.6)			
		Republic of Ireland	217 (0.7)	162 (0.7)	55 (0.6)			
		Scotland	2296 (7.1)	1693 (7.1)	603 (7.1)			
		Wales	600 (1.8)	425 (1.8)	175 (2.1)			
		other	13 (0.0)	10 (0.0)	3 (0.0)			
	Current employment status - visit 2, n (%)	In paid employment or self-employed	12978 (39.9)	10528 (43.9)	2450 (28.7)	<0.001	Chi-squared	0.338
		Retired	18302 (56.3)	12469 (52.0)	5833 (68.3)			
		other	1218 (3.7)	965 (4.0)	253 (3.0)			
	Current smoker, n (%)	0	31893 (98.1)	23497 (98.1)	8396 (98.4)	0.086	Chi-squared	0.023
		1	605 (1.9)	465 (1.9)	140 (1.6)			
	Current tobacco smoking - visit 2, n (%)	No	31431 (96.7)	23125 (96.5)	8306 (97.3)	0.001	Chi-squared	0.050
		Yes, on most or all days	605 (1.9)	465 (1.9)	140 (1.6)			
		other	462 (1.4)	372 (1.6)	90 (1.1)			
	Diabetic, n (%)	0	31154 (95.9)	23462 (97.9)	7692 (90.1)	<0.001	Chi-squared	0.333
		1	1344 (4.1)	500 (2.1)	844 (9.9)			
	Ethnic background - visit 2, n (%)	British	7408 (22.8)	5305 (22.1)	2103 (24.6)	<0.001	Chi-squared	0.059
		other	25090 (77.2)	18657 (77.9)	6433 (75.4)			
	Ever smoked - visit 2, n (%)	No	15217 (46.8)	11408 (47.6)	3809 (44.6)	<0.001	Chi-squared	0.060
		Yes	17196 (52.9)	12495 (52.1)	4701 (55.1)			
		other	85 (0.3)	59 (0.2)	26 (0.3)			
	Handedness (chirality/laterality) - visit 0, n (%)	Left-handed	3043 (9.4)	2242 (9.4)	801 (9.4)	0.406	Chi-squared (warning: expected count < 5)	0.021
		Right-handed	28957 (89.1)	21369 (89.2)	7588 (88.9)			
		Use both right and left hands equally	486 (1.5)	342 (1.4)	144 (1.7)			
		other	12 (0.0)	9 (0.0)	3 (0.0)			
	Hypercholesterol, n (%)	0	24538 (75.5)	20377 (85.0)	4161 (48.7)	<0.001	Chi-squared	0.836
		1	7960 (24.5)	3585 (15.0)	4375 (51.3)			
	Inflammation, n (%)	False	27267 (83.9)	20553 (85.8)	6714 (78.7)	<0.001	Chi-squared	0.187
		True	5231 (16.1)	3409 (14.2)	1822 (21.3)			
	Job involves shift work - visit 2, n (%)	Always	795 (2.4)	644 (2.7)	151 (1.8)	<0.001	Chi-squared	0.318
		Never/rarely	11077 (34.1)	9008 (37.6)	2069 (24.2)			
		Sometimes	855 (2.6)	678 (2.8)	177 (2.1)			
		other	19771 (60.8)	13632 (56.9)	6139 (71.9)			
	Length of time at current address - visit 2, mean (SD)		21.0 (13.9)	20.4 (13.5)	22.7 (14.6)	<0.001	Two Sample T-test	0.166
	Life long depression, n (%)	False	27461 (84.5)	20138 (84.0)	7323 (85.8)	<0.001	Chi-squared	0.049
		True	5037 (15.5)	3824 (16.0)	1213 (14.2)			
	Ls pc dim.1, mean (SD)		0.6 (2.0)	0.7 (2.0)	0.4 (2.1)	<0.001	Two Sample T-test	-0.169
	Ls pc dim.2, mean (SD)		-0.4 (2.0)	-0.5 (2.0)	-0.4 (2.0)	0.009	Two Sample T-test	0.033
	Maternal smoking around birth - visit 0, n (%)	Do not know	3388 (10.4)	2419 (10.1)	969 (11.4)	<0.001	Chi-squared (warning: expected count < 5)	0.072
		No	20344 (62.6)	15217 (63.5)	5127 (60.1)			
		Yes	8749 (26.9)	6313 (26.3)	2436 (28.5)			
		other	17 (0.1)	13 (0.1)	4 (0.0)			
	Mh pc dim.1, mean (SD)		1.0 (6.5)	1.1 (6.5)	0.8 (6.7)	0.004	Two Sample T-test	-0.037
	Mh pc dim.2, mean (SD)		0.5 (3.3)	0.5 (3.3)	0.4 (3.3)	0.019	Two Sample T-test	-0.029
	Number in household - visit 2, mean (SD)		2.2 (1.0)	2.2 (1.1)	2.0 (0.8)	<0.001	Two Sample T-test	-0.161
	Number of full brothers - visit 2, mean (SD)		1.0 (1.1)	1.1 (1.1)	1.0 (1.1)	0.011	Two Sample T-test	-0.032

...continued

variable role	variable name	variable modalities	Overall	No hypertension	Hypertension	P-Value	Test	SMD (0,1)
	Number of full sisters - visit 2, mean (SD)		1.0 (1.1)	1.0 (1.1)	1.0 (1.1)	0.365	Two Sample T-test	-0.012
	Number of vehicles in household - visit 2, n (%)	Four or more	922 (2.8)	729 (3.0)	193 (2.3)	<0.001	Chi-squared	0.118
		None	1265 (3.9)	902 (3.8)	363 (4.3)			
		One	13770 (42.4)	9885 (41.3)	3885 (45.5)			
		Three	3034 (9.3)	2380 (9.9)	654 (7.7)			
		Two	13471 (41.5)	10037 (41.9)	3434 (40.2)			
		other	36 (0.1)	29 (0.1)	7 (0.1)			
	Obesity, n (%)	0	26781 (82.4)	20681 (86.3)	6100 (71.5)	<0.001	Chi-squared	0.370
		1	5717 (17.6)	3281 (13.7)	2436 (28.5)			
	Part of a multiple birth - visit 0, n (%)	No	31809 (97.9)	23448 (97.9)	8361 (97.9)	0.242	Chi-squared	0.021
		Yes	644 (2.0)	485 (2.0)	159 (1.9)			
		other	45 (0.1)	29 (0.1)	16 (0.2)			
	Past tobacco smoking - visit 2, n (%)	I have never smoked	15218 (46.8)	11409 (47.6)	3809 (44.6)	<0.001	Chi-squared	0.181
		Just tried once or twice	5296 (16.3)	4118 (17.2)	1178 (13.8)			
		Smoked occasionally	4008 (12.3)	3001 (12.5)	1007 (11.8)			
		Smoked on most or all days	7287 (22.4)	4911 (20.5)	2376 (27.8)			
		other	689 (2.1)	523 (2.2)	166 (1.9)			
	Qualifications - visit 2, n (%)	A levels/AS levels or equivalent	4013 (12.3)	3024 (12.6)	989 (11.6)	<0.001	Chi-squared	0.199
		CSEs or equivalent	1205 (3.7)	901 (3.8)	304 (3.6)			
		College or University degree	16203 (49.9)	12376 (51.6)	3827 (44.8)			
		NVQ or HND or HNC or equivalent	1901 (5.8)	1261 (5.3)	640 (7.5)			
		None of the above	1762 (5.4)	1098 (4.6)	664 (7.8)			
		O levels/GCSEs or equivalent	5717 (17.6)	4128 (17.2)	1589 (18.6)			
		Other professional qualifications eg: nursing, teaching	1596 (4.9)	1111 (4.6)	485 (5.7)			
		other	101 (0.3)	63 (0.3)	38 (0.4)			
	Scanner lateral (x) brain position - visit 2, mean (SD)		0.7 (3.0)	0.7 (3.0)	0.8 (3.0)	<0.001	Two Sample T-test	0.051
	Scanner longitudinal (z) brain position - visit 2, mean (SD)		-28.1 (26.8)	-27.5 (26.4)	-29.8 (27.7)	<0.001	Two Sample T-test	-0.086
	Scanner table position - visit 2, mean (SD)		-1054.2 (25.1)	-1053.9 (24.8)	-1055.1 (26.0)	<0.001	Two Sample T-test	-0.045
	Scanner transverse (y) brain position - visit 2, mean (SD)		66.4 (5.9)	66.2 (5.9)	66.9 (6.0)	<0.001	Two Sample T-test	0.109
	Self declared mental exclude visit 2, mean (SD)		0.0 (0.2)	0.0 (0.1)	0.0 (0.2)	<0.001	Two Sample T-test	0.120
	Sex - visit 0, n (%)	Female	17195 (52.9)	13679 (57.1)	3516 (41.2)	<0.001	Chi-squared	0.322
		Male	15303 (47.1)	10283 (42.9)	5020 (58.8)			
	Smoking status - visit 2, n (%)	Current	1065 (3.3)	835 (3.5)	230 (2.7)	<0.001	Chi-squared	0.153
		Never	20503 (63.1)	15521 (64.8)	4982 (58.4)			
		Previous	10846 (33.4)	7548 (31.5)	3298 (38.6)			
		other	84 (0.3)	58 (0.2)	26 (0.3)			
	Social support, n (%)	False	12720 (39.1)	9175 (38.3)	3545 (41.5)	<0.001	Chi-squared	0.066
		True	19778 (60.9)	14787 (61.7)	4991 (58.5)			
	Total alcohol weekly units 2, mean (SD)		13.1 (14.4)	12.6 (13.6)	14.5 (16.4)	<0.001	Two Sample T-test	0.125
	Townsend deprivation index at recruitment - visit 0, mean (SD)		-1.9 (2.7)	-1.9 (2.7)	-1.9 (2.7)	0.435	Two Sample T-test	-0.010
	Uk biobank assessment centre - visit 2, n (%)	Cheadle	18991 (58.4)	13954 (58.2)	5037 (59.0)	0.241	Chi-squared	0.026
		Newcastle	8436 (26.0)	6215 (25.9)	2221 (26.0)			
		Reading	5028 (15.5)	3763 (15.7)	1265 (14.8)			
		other	43 (0.1)	30 (0.1)	13 (0.2)			
	Volumetric scaling from t1 head image to standard space - visit 2, mean (SD)		1.3 (0.1)	1.3 (0.1)	1.3 (0.1)	<0.001	Two Sample T-test	-0.169

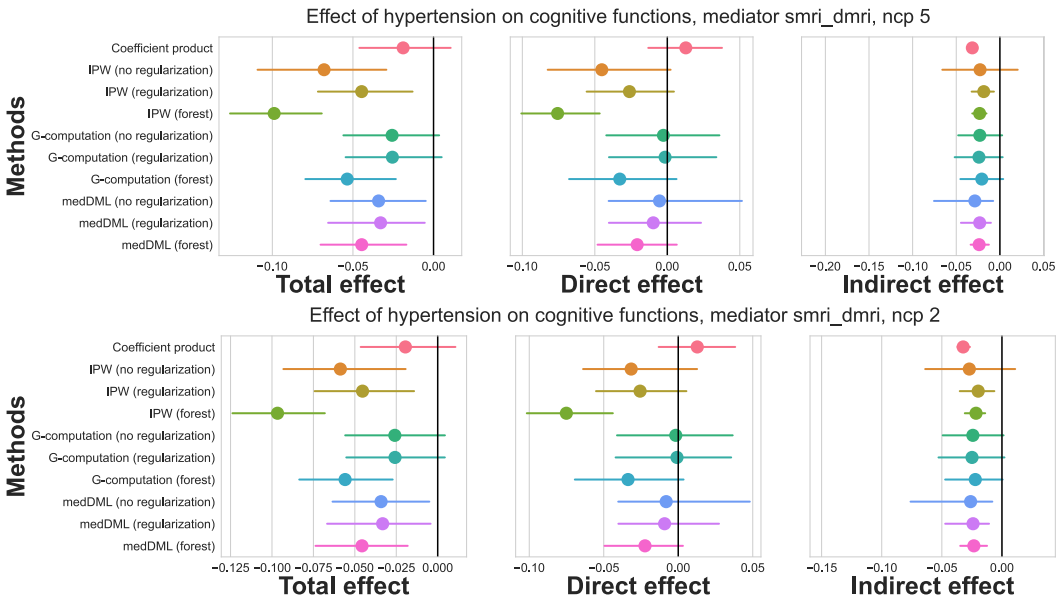


Figure S19: **Causal total, direct and indirect effect of hypertension on cognitive functions, mediated through the brain structure.** The top panel represents results presented in the main text, with 5 components for the imputation of missing values by PCA for confounding variables, and the bottom panel with 2 components. We observe very similar results. We also show the estimates obtained with different estimators, and again, results are quite similar among the different approaches. Each estimate is represented as a mean over 100 bootstrapped samples. The error bar is the 95% percentile-interval, estimated from the empirical bootstrap distribution.

Listing 1: Dagitty code to reproduce the causal graph used for the mediation analysis.

```
dag {
"brain MRI" [pos="-0.218,1.511"]
"brain structure" [pos="-0.575,0.931"]
"cognitive functions" [outcome,pos="1.301,1.740"]
"college degree" [pos="-1.479,0.467"]
"current depression" [pos="0.573,0.518"]
"current smoking" [pos="-0.064,-0.580"]
"drug use" [pos="-1.166,-0.731"]
"early life factors" [latent,pos="-0.856,-1.068"]
"life long depression" [pos="0.337,-0.074"]
"lifestyle factors" [latent,pos="-1.608,-0.454"]
"maternal smoking" [pos="-1.694,-1.236"]
"number of vehicles" [pos="0.451,-1.387"]
"physical activity" [pos="-2.063,-0.610"]
"screen use" [pos="-1.617,-0.903"]
"social support" [pos="0.967,0.141"]
"socio economic factors" [latent,pos="-0.027,-1.257"]
"townsend index" [pos="0.615,-1.730"]
"weight and height age 10" [pos="-1.640,-1.805"]
U [latent,pos="-2.200,-1.520"]
age [pos="0.954,-1.351"]
alcohol [pos="-1.846,0.848"]
diabetes [pos="-1.041,0.019"]
hypercholesterol [pos="-0.623,0.082"]
hypertension [exposure,pos="-1.987,1.581"]
inflammation [pos="0.863,-0.259"]
obesity [pos="1.414,0.370"]
revenue [pos="-0.575,-1.590"]
sex [pos="1.410,-1.112"]
siblings [pos="-0.496,-1.213"]
```

```
"brain structure" -> "brain MRI"
"brain structure" -> "cognitive functions"
"college degree" -> "brain structure"
"college degree" -> "cognitive functions"
"current depression" -> "brain structure"
"current depression" -> "cognitive functions"
"current smoking" -> "brain structure"
"current smoking" -> "cognitive functions"
"early life factors" -> "brain structure"
"early life factors" -> "cognitive functions"
"early life factors" -> "college degree"
"early life factors" -> "current depression"
"early life factors" -> "current smoking"
"early life factors" -> "life long depression"
"early life factors" -> "maternal smoking"
"early life factors" -> "social support"
"early life factors" -> "weight and height age 10"
"early life factors" -> alcohol
"early life factors" -> diabetes
"early life factors" -> hypercholesterol
"early life factors" -> hypertension
"early life factors" -> inflammation
"early life factors" -> obesity
"early life factors" -> siblings
"life long depression" -> "brain structure"
"life long depression" -> "cognitive functions"
"lifestyle factors" -> "brain structure"
"lifestyle factors" -> "cognitive functions"
"lifestyle factors" -> "college degree"
"lifestyle factors" -> "current depression"
"lifestyle factors" -> "current smoking"
"lifestyle factors" -> "drug use"
"lifestyle factors" -> "life long depression"
"lifestyle factors" -> "physical activity"
"lifestyle factors" -> "screen use"
"lifestyle factors" -> "social support"
"lifestyle factors" -> alcohol
"lifestyle factors" -> diabetes
"lifestyle factors" -> hypercholesterol
"lifestyle factors" -> hypertension
"lifestyle factors" -> inflammation
"lifestyle factors" -> obesity
"social support" -> "brain structure"
"social support" -> "cognitive functions"
"socio economic factors" -> "brain structure"
"socio economic factors" -> "cognitive functions"
"socio economic factors" -> "college degree"
"socio economic factors" -> "current depression"
"socio economic factors" -> "current smoking"
"socio economic factors" -> "life long depression"
"socio economic factors" -> "number of vehicles"
"socio economic factors" -> "social support"
"socio economic factors" -> "townsend index"
"socio economic factors" -> alcohol
"socio economic factors" -> diabetes
"socio economic factors" -> hypercholesterol
"socio economic factors" -> hypertension
"socio economic factors" -> inflammation
"socio economic factors" -> obesity
"socio economic factors" -> revenue
U -> "current smoking"
U -> "early life factors"
U -> "lifestyle factors"
U -> "socio economic factors"
age -> "brain structure"
```

```
age -> "cognitive functions"
age -> "college degree"
age -> "current depression"
age -> "current smoking"
age -> "life long depression"
age -> "social support"
age -> alcohol
age -> diabetes
age -> hypercholesterol
age -> hypertension
age -> inflammation
age -> obesity
alcohol -> "brain structure"
alcohol -> "cognitive functions"
diabetes -> "brain structure"
diabetes -> "cognitive functions"
hypercholesterol -> "brain structure"
hypercholesterol -> "cognitive functions"
hypertension -> "brain structure"
hypertension -> "cognitive functions"
inflammation -> "brain structure"
inflammation -> "cognitive functions"
obesity -> "brain structure"
obesity -> "cognitive functions"
sex -> "brain structure"
sex -> "cognitive functions"
sex -> "college degree"
sex -> "current depression"
sex -> "current smoking"
sex -> "life long depression"
sex -> "social support"
sex -> alcohol
sex -> diabetes
sex -> hypercholesterol
sex -> hypertension
sex -> inflammation
sex -> obesity
}
```