Supplementary Materials of "Dissecting the colocalized GWAS and eQTLs with mediation analysis for high dimensional exposures and confounders"

Qi Zhang
Department of Mathematics and Statistics,
University of New Hampshire
qi.zhang2@unh.edu

Zhikai Yang
PhD student in Complex Biosystems Program
and
Department of Agronomy and Horticulture,
University of Nebraska-Lincoln

Jinliang Yang
Department of Agronomy and Horticulture,
University of Nebraska-Lincoln

November 13, 2023

Web Appendix A: the equivalence of the difference-in-coefficients approach and the product-of-coefficients approach for least square estimates

Without losing generalities, we consider the following linear meidation models.

$$\begin{aligned} Y &= Z\tilde{\beta} + \tilde{\epsilon} \\ Y &= Z\beta + M\gamma + \epsilon \\ M &= ZB + \eta \end{aligned}$$

Here Z is the matrix of the exposures and confounders combined. If there are q exposures and s confounders, Z is a matrix with $q+s$ columns, and we are only interested in the mediation effects for the exposures represented by the first q columns of Z. We assume that q, s and p, the number of mediators, are all low-dimensional so that the least square estimates of the individual equations work well. Specifically, the least square estimates are

$$\begin{aligned} \hat{\tilde{\beta}} = &\quad (Z^T Z)^{-1} Z^T Y \\ \hat{B} = &\quad (Z^T Z)^{-1} Z^T M \\ \hat{\beta} = &\quad (C_{11} Z^T + C_{12} M^T) Y \\ \hat{\gamma} = &\quad (C_{12}^T Z^T + C_{22} M^T) Y \end{aligned}$$

where

$$\begin{aligned} C_{22} = &\quad (M^T M - M^T Z (Z^T Z)^{-1} Z^T M)^{-1} \\ C_{11} = &\quad (Z^T Z)^{-1} + (Z^T Z)^{-1} Z^T M C_{22} M^T Z (Z^T Z)^{-1} \\ C_{12} = &\quad -(Z^T Z)^{-1} Z^T M C_{22} \end{aligned}$$

The following derivation shows that the difference-in-coefficients estimator based on the least squares is equivalent to the estimator based on the product-of-coefficients approach.

$$
\begin{aligned}
\hat{\tilde{\beta}} - \hat{\beta} &= (Z^T Z)^{-1} Z^T Y - (C_{11} Z^T + C_{12} M^T) Y \\
&= \left\{ (Z^T Z)^{-1} Z^T - (Z^T Z)^{-1} Z^T - (Z^T Z)^{-1} Z^T M C_{22} M^T Z (Z^T Z)^{-1} Z^T + (Z^T Z)^{-1} Z^T M C_{22} M^T \right\} Y \\
&= (Z^T Z)^{-1} Z^T M \left\{ C_{22} M^T - C_{22} M^T Z (Z^T Z)^{-1} Z^T \right\} Y \\
&= (Z^T Z)^{-1} Z^T M \left\{ C_{22} M^T + C_{12}^T Z^T \right\} Y \\
&= \hat{B} \hat{\gamma}
\end{aligned}
$$

This result is reassuring, and enables us to compare the two estimation approaches for the same target parameter. As illustrated in [15] and our study, different penalized estimators yield different asymptotic properties and empirical performance.

Web Appendix B: Implementation of MedDiC

Most of the current implementations of the debiased lasso and the scaled lasso are slow or do not accommodate our mediation problem directly [2, 10, 12]. Thus, we implement MedDiC from scratch using RCPP, leveraging some existing algorithms scattered in various packages such as fastGGM [11] and hdi [2].

Algorithm 1 Scaled adaptive lasso for linear model $y \sim N(X\beta, \sigma^2 I_n)$

Input: $n \times p$ covariate matrix X and length n outcome vector y

Output: $\hat{\beta}$ and $\hat{\sigma}^2$

1: *Initial estimate of β:* Apply the scaled lasso [10] with penalty term $\sum_{j=1}^{p} |\beta_j|$ to regress y on X, and return $\hat{\beta}^{(0)}$, the estimated coefficients.

2: *Construct weights for scaled adaptive lasso:* Let $w = (w_1, \ldots, w_p) \propto 1/(|\hat{\beta}^{(0)}| + \sqrt{var(y)/n})$ such that $\sum_{j=1}^{p} w_j = p$.

3: *Final estimate of β:* Minimize the scaled lasso objective with penalty term $\sum_{j=1}^{p} w_j |\beta_j|$ instead of $\sum_{j=1}^{p} |\beta_j|$, and return $\hat{\beta}$.

4: *Estimate σ^2:* Applied least square to regress y to the selected covariates based on $\hat{\beta}$, and return the mean squared error as $\hat{\sigma}^2$.

We uses a scaled adaptive lasso (Algorithm 1) as the initial estimator for the debiased lasso

to reduce bias. This algorithm constructs the weights using estimated coefficients from an initial fit, a strategy also used in previous studies [14, 16]. We estimate the noise level based on the least square after the scaled lasso, because [9] shows that it is less biased than using the direct output of the original scaled lasso. Although [9] advocates for cross-validation for noise level estimation and [2] suggests using lasso with cross-validation for the initial estimator of the debiased lasso, our exploratory simulation experiments show no clear gain in high dimensional setting to justify the additional computational cost of cross-validation. The score calculations are also based on the scaled lasso. The tuning parameter of the scaled lasso is $\sqrt{2\log(s+q)/n}$ for the models without the mediators, and $\sqrt{2\log(s+q+p)/n}$ for models with mediators. For the low dimensional problems with $q+s < c \cdot n$ where $c \in (0,1)$ is a fixed constant, the ordinary least square is used instead of the debiased lasso for estimating $\tilde{\beta}^\star$. In this study, we set $c = 0.2$.

The proposed MedDiC procedure is summarized as Algrorithm 2 where the equation numbers are referred to the equations in the main manuscript.

Algorithm 2 MedDiC: Mediation analysis using Difference in Coefficients

Input: The length n response vector Y, the $n \times p$ mediator matrix M, $Z^\star = (X, Z)$ where X is the $n \times s$ confounder matrix and Z is the $n \times q$ exposure matrix.

Output: $\hat{\tilde{\beta}}^\star - \hat{\beta}^\star$ the estimated indirect effect, and the standard errors.

1: *Estimate the total effect $\tilde{\beta}^\star$:* Apply Equation (6) where $\tilde{R}$ and $\tilde{W}$ are calculated according to [12], and $\hat{\tilde{\beta}}^{\star,init}$ is based Algorithm 1, and returns $\hat{\tilde{\beta}}^\star$.

2: *Estimate the direct effect $\beta^\star$:* Apply Equation (8) where R and W are calculated according to [12], and $\hat{\beta}^{\star,init}$ is based on Algorithm 1, and returns $\hat{\beta}^\star$.

3: *Estimate the indirect effect $\tilde{\beta}^\star - \beta^\star$:* returns $\hat{\tilde{\beta}}^\star - \hat{\beta}^\star$

4: *Estimate the standard error of indirect effect:* For $j = 1+s, \ldots, q+s$ calculate the standard errors of $\hat{\tilde{\beta}}_j^\star - \hat{\beta}_j^\star$ based on Equation (10).

Web Appendix C: Simulation model setup

We generate simulated data using the outcome model and the mediator model

$$Y = X\alpha + Z\beta + M\gamma + \epsilon \quad (1)$$

$$M = XA + ZB + \eta \quad (2)$$

For $i = 1, \ldots, n$, we simulate $X_{i,:} = (X_{i1}, \ldots, X_{is}) \sim N(0, I_s)$, $Z_{i,:} = (Z_{i1}, \ldots, Z_{iq}) \sim N(0, \Sigma)$ where $\Sigma = \left(r^{|j-k|}\right)_{j,k=1}^{q}$ is a toeplitz matrix with entries $\Sigma_{jk} = r^{|j-k|}$. The cases with $r \neq 0$ represents the scenarios in which the multivariate exposures are correlated through unmeasured confounders. We also simulate $\eta_i \sim N(0, I_p)$ and $\epsilon_i \sim N(0, 1)$. We set $s = 2$, $\alpha = (-0.2, 0.2)^T$, and A to be a matrix whose rows are random permutations of $\{-0.1, 0.1)\}$.

Recall that the indirect and the direct effect of exposure j are $B_{j,:}\gamma$, and β_j, respectively, where $B_{j,:}$ is the jth row of matrix B. We simulate β, B and γ such that only four out of q exposures have non-zero indirect and/or direct effects. The effect sizes $(B_{j,:}\gamma,\ \beta_j)$ for these four exposures are chosen to represent exposures with different effects: $(-\tau, 0)$ for only indirect effect (complete mediation), $(0, -\tau)$ for only direct effect, (τ, τ) for both effects with the same sign, and $(1.2\tau, -0.8\tau)$ for both effects with different signs. Here τ is the simulation parameter for the signal strength.

In the following, we describe in more details how γ and B are simulated to satisfy the above requirement. Let p_m be the number of true mediators. We first simulate γ as a sparse vector with $2p_m$ non-zero elements with value $\sqrt{\tau}$. This choice is made so that the non-zero elements in both B and γ grow as the signal strength τ increases. Half of these $2p_m$ candidate mediators are true mediators, and the other half are referred to as "spurious mediators". They have non-zero coefficients in the outcome model, but zero link to the exposures. Let $C_{true}, C_{spur} \subset \{1, \ldots, p\}$ be their coordinate sets, and $S \subset \{1, \ldots, q\}$ be the subset of exposures with indirect effects. We initialize the $q \times p$ matrix B with 10 nonzero elements at random locations in each row. These non-zero elements are random samples from a uniform distribution between -1 and 1. We then modify B as the following. $B_{S^c, C_{true}}$ is for the association between the exposures with no indirect effect and the true mediators, and $B_{:, C_{spur}}$ is for the association between all exposures and the fake mediators. By definitions of S^c and C_{spur}, their elements are replaced with zeros. $B_{S, C_{true}}$ is the sub-matrix of the association between the exposures with non-zero mediation effects and the

true mediators. We replace the elements in this sub-matrix row-wise with repeating sequence of 1,0,2, allowing each exposure with an indirect effect to influence a distinct subset of true mediators. When $p_m = 1$, however, all exposures with non-zero mediation effect must influence this only true mediator, and $B_{S,C_{true}}$ is filled with the sequence of 1,1,2 instead. We further normalize B row-wise so that the nonzero elements of $B\gamma$ are $-\tau, \tau$, and 1.2τ. We then simulate β as a sparse vector with three nonzero elements $-\tau, \tau, -0.8\tau$. Finally, vectors $B\gamma$ and β are jointly shuffled such that there are only four exposures that have nonzero direct and/or indirect effect, and their values are as specified in the last paragraph.

In this simulation study, we set $n = 300$, $p = 500$, $s = 2$, and explore various signal strength $\tau \in \{0.25, 0.5, 0.75, 1, 1.5, 2\}$. Our proposed method is evaluated for low dimensional exposures $q = 5$, and high dimensional exposures $q = 400$, with consideration given to both uncorrelated exposures ($r = 0$), and correlated exposures ($r = 0.4$). For low dimensional exposures, we examine both $p_m = 1$ and $p_m = 5$, while for high dimensional exposures, we only consider $p_m = 5$. We repeat each simulation setting for 100 times, except for the high dimensional adaptations of [15], which are based on 40 replicates due to their high computational cost. The results for the two versions of mediateR [6] are based on 20 replicates due to the high computational cost and poor performance. The testing error rates, powers, coverage probabilities and the width distribution of the confidence intervals are calculated over all exposures in all simulation replicates. For instance, in each of the 100 simulation replicate, there are three exposures with non-zero indirect effect, so the empirical power of detecting the indirect effect is the proportion detected among the 3×100 true non-zero indirect effects.

Web Appendix D: Methods to be compared in simulations and real data analysis

MedDiC is designed for high dimensional exposures, and it can also be applied to low dimensional exposures. We directly compare MedDiC with two state-of-art methods for the inference of indirect effects with low dimensional exposures. They are ZWZ [15] and GLLZ [3]. As the original code of [15] did not report any results for the inference of direct effect, and that of [3] did not provide any confidence intervals, we modified their code to include these results based on

their asymptotic distributions. We compare these methods in terms of empirical FDR, power, and the marginal coverage probability and widths of 95% confidence intervals for low dimensional exposures.

When the exposures are high dimensional, GLLZ and ZWZ cannot be directly applied, and there is no method for this setting except the proposed MedDiC method and mediateR. For benchmarking purpose, we adapt each of GLLZ and ZWZ in the following two different ways for high dimensional exposures. This results in four methods to be compared with the proposed MedDiC method in high dimensional settings.

(1) After Marginal Screening (**GLLZ AMS** and **ZWZ AMS**): We first apply screening by marginal correlation between the exposures and the outcome, keep the top 20, and then apply GLLZ or ZWZ to the survived exposures. The exposures that are not used in the final model fit return p-value= 1, and a confidence interval $[0, 0]$. Assuming a true effect of 0, we consider this interval to cover the true value in our evaluations.

(2) Univariate Scanning (**GLLZ US** and **ZWZ US**): This is essentially what [15] has used in their real data analysis. Let X_0 be the low dimensional confounder matrix, and U be the top left singular vectors of the high dimensional exposure matrix Z (top five in simulations, and top 20 in real data analysis). For $j = 1, \ldots, q$, GLLZ or ZWZ is applied to exposure matrix $(X_0, U, Z_{:,j})$. It requires to run GLLZ or ZWZ q times for a dataset. We apply GLLZ US to real data analysis, while ZWZ US is excluded due to computational constraints. For simulation studies, we further modify ZWZ US to reduce the computational cost. In our simulations, we only run ZWZ US to a subset of exposures $S_o \subset \{1, \ldots, q\}$ that include all non-zero effects. This subset is chosen based on the simulation model as the following. If $j \in \{1, \ldots, q\}$ is an exposure with non-zero indirect or direct effect, then $\{j-1, j, j+1\} \cap \{1, \ldots, q\} \subset S_0$. Since there are four exposures with non-zero effects in the simulation model, ZWZ US only need to run up to 12 times, rather than 400 times for each simulation replicate. For exposures $j \notin S_o$, ZWZ US return p-value= 1, and $[0, 0]$ as confidence interval. We remark that all true signals are used in fitting the ZWZ model, which is different from ZWZ AMS. Since the computational cost of ZWZ increases roughly linearly with q, this modification allows us to include ZWZ US in simulation-based comparisons at 3% of the computational cost without losing any true signals.

HIMA[13] is one of the most widely recognized mediator selection algorithm, and it has been

recently modified as HIMA2 [8]. Their statistical inference objective differs from ours, and cannot be compared with MedDiC, ZWZ or GLLZ in general settings. As discussed in [15], HIMA's objective aligns with ours when there is only one true mediator. For the low dimensional exposure ($q = 5$) and when $p_m = 1$, we include HIMA and HIMA2 for comparing FDR and power. They are designed for univariate exposure. Thus we apply them to each exposure separately. It should be noted that it requires q separate runs for each simulation replicate. For each exposure, we use the p-value of its most significant mediator as the p-value of the overall mediation effect of this exposure.

The mediateR package [6] proposed a unified inference framework for mediation analysis based on the product-of-coefficients approach. Its estimation is based on penalized generalized linear model and the inference is based on nonparametric bootstrap. In the original paper, this framework was only evaluated for the binary and survival outcomes using ridge penalty. This conceptual framework can be applied the settings with high dimensional mediators, exposures and confounders. We refer to mediateR model with ridge penalty as medR L2 method. We also extend the implementation of mediateR to utilize the lasso penalty for better computational efficiency, and refer to it as medR L1. To accommodate the sparsity penalty, the definition of p-value for mediateR (the paragraph after equation (7) on page 5 of [6]) is modified as $2min(P_L, P_U) + \frac{1}{B}\sum_{b=1}^{B} 1_{\hat{IE}_{X_i}^{(b)}=\Delta}$. Without such modification, the bootstrap replicates whose estimated coefficient is exactly 0 are not counted towards the p-value when $\Delta = 0$. For one simulation setting $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$, we compare their performance and computing time with 500 bootstrap replicates. We decide to exclude MedR L1 and MedR L2 from further simulation studies due to their poor performance (Web Figures 1,7,13) and the high computational cost (See Web Appendix F for details). In particular, if fully included in our simulation studies, the estimated computing times for MedR L1 and MedR L2 are of the order of 10^4 hours and 10^5 hours, respectively . medR L1 is included in the two real data analysis.

To account for multiple testing, we apply Benjamini-Hochberg procedure[1] at level $\alpha = 0.05$ to all methods. At this level, we compare the empirical FDR and empirical power as defined in [14, 2, 13]. We also compare the empirical marginal coverage probability and wdiths for 95% confidence intervals.We plot the marginal coverage probabilities of the confidence intervals for exposures with zero and non-zero effects in separate figures, because their behaviors are different

in some settings.

We recap the methods included in the comparison for each analysis. For the simulations with low dimensional exposures and only one true mediator, we compared MedDiC with ZWZ, GLLZ, HIMA and HIMA2. For the one setting $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$ that has been used for computing time comparisons, MedR L1 and MedR L2 are also included in the comparisons with 20 replicates. For the simulations with low dimensional exposures and five true mediators, we compared MedDiC with ZWZ, GLLZ. For the simulations with high dimensional exposures, we compared MedDiC with ZWZ AMS, ZWZ US, GLLZ AMS, and GLLZ US. The only exception is the comparison in the width of the confidence intervals, ZWZ AMS, GLLZ AMS and ZWZ US are excluded because they return intervals with 0 widths for the screened exposures. For the analysis of the Maize data where the exposure is high dimensional and the number of mediators is slightly smaller than the sample size, MedDiC was compared with GLLZ US and MedR L1. For the analysis of the Mice data with high dimensional mediators, high dimensional confounders, and the number of exposures is about one half of the sample size, MedDiC was compared with GLLZ, ZWZ and MedR L1.

Web Appendix E: Simulation results for the inference of direct effects and total effects

The primary focus of this paper is to infer indirect effect, and we do not claim that inferring the direct and total effects from MedDiC is a novel contribution. This is because they are the direct outputs of the debiased lasso.

Despite this, we present the inference results of direct effects from MedDiC, and compare them with mediateR, ZWZ and GLLZ (and their high dimensional adaptations) in Web Figures 7-12. For low dimensional exposures, MedDiC performs comparably to ZWZ and GLLZ, and the empirical coverage probabilities of non-zero direct effects for all three methods are slightly lower than the nominal level. MedR L1 and MedR L2 have poor coverage probability for the nonzero direct effects. Additionally, MedR L2 also yields poor FDR control and the coverage probability for the zero direct effect. For high dimensional exposures, we find that MedDiC and GLLZ US show the highest power. However, GLLZ US fails to control FDR near the nominal level. The empirical coverage probabilities of non-zero direct effects for all methods are lower than the nominal level, and GLLZ AMS and ZWZ AMS still perform the worst based on this measure. MedDiC yields shorter confidence intervals than GLLZ US.

Web Figure 13-18 reports the inference results of total effects from MedDiC. As noted earlier, they are the direct outputs of the debiased lasso. We do not claim this part as a novel contribution, and only present the simulation results for completeness. ZWZ and GLLZ do not report or estimate the total effects. We find that the inference of the total effect is valid in all scenarios in terms of FDR control and coverage probabilities of the confidence intervals, and the power is reasonable. MedR L1 and MedR L2 return low coverage probability for the nonzero total effects.

Web Appendix F: Simulation results for computational time comparison

We also evaluate the computational cost of MedDiC and compare with ZWZ, GLLZ, HIMA, HIMA2, MedR L1 and MedRL2 on the same computer. This part of the simulation were per-

formed on a laptop with Intel(R) i7-1065G7 CPU @1.30GHz 1.50GHz, 32 GB memory, Win 10 operation system and Microsoft R Open 4.02 for computing. We remark that Microsoft R Open 4.02 is capable of utilizing multiple cores. In Web Table 1, we present the average computational time (in seconds) of MedDiC, ZWZ, HIMA, HIMA2, GLLZ, MedR L1 and MedR L2 in the simulation setting with $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$. We find that MedDiC and GLLZ are much faster than the rest. GLLZ is faster because it does not need to calculate the score matrix for the debiased lasso. We find that our extension of the mediateR method (MedR L1) does reduce the computational cost comparing with the the MedR L2 as implemented in [6]. However, MedR L1 and MedR L2 remain too time consuming to be fully included in all simulation settings. The estimated pure computing times for MedR L1 and MedR L2, if fully included in all simulation settings with 500 bootstrap replicates, are at the order of 10^4 hours and 10^5 hours using the same machine. For the high dimensional setting, We only compare MedDiC and the two high dimensional adaptations of GLLZ, and we find that GLLZ AMS is fast due to the screening step, and GLLZ US is more than 10 times slower than MedDiC because it takes $q = 400$ GLLZ runs. We do not include the other methods, because they are expected to be much slower than in the low dimensional setting as more exposures are included in the model fitting.

We remark that we do not document the computational cost for all our simulation settings or using more replicates largely due to the high computational cost of some of the competing methods. We use parallel computing and multiple computers, and it is difficult to compare the computational times consistently in this setting.

Web Appendix G: Generating the list of known TFs for mouse

To gather a list of known transcription factors for *Mus musculus* , we collect data from three sources [4, 5, 7]. Each list contains over 1300 TFs, with a merged list of 1929 TFs.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[2] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

[3] X. Guo, R. Li, J. Liu, and M. Zeng. High-dimensional mediation analysis for selecting dna methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association*, 117(539):1110–1121, 2022.

[4] J. Hammelman, T. Patel, M. Closser, H. Wichterle, and D. Gifford. Ranking reprogramming factors for cell differentiation. *Nature Methods*, 19(7):812–822, 2022.

[5] H. Hu, Y.-R. Miao, L.-H. Jia, Q.-Y. Yu, Q. Zhang, and A.-Y. Guo. Animaltfdb 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research*, 47(D1):D33–D38, 2019.

[6] L. Huang, J. P. Long, E. Irajizad, J. D. Doecke, K.-A. Do, and M. J. Ha. A unified mediation analysis framework for integrative cancer proteogenomics with clinical outcomes. *Bioinformatics*, 39(1):btad023, 2023.

[7] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, et al. Gateways to the fantom5 promoter level mammalian expression atlas. *Genome biology*, 16(1):1–14, 2015.

[8] C. Perera, H. Zhang, Y. Zheng, L. Hou, A. Qu, C. Zheng, K. Xie, and L. Liu. Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. *BMC bioinformatics*, 23(1):1–14, 2022.

[9] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

[10] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[11] T. Wang, Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang, and W. Chen. Fastggm: an efficient algorithm for the inference of gaussian graphical model in biological networks. *PLoS computational biology*, 12(2):e1004755, 2016.

[12] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[13] H. Zhang, Y. Zheng, Z. Zhang, T. Gao, B. Joyce, G. Yoon, W. Zhang, J. Schwartz, A. Just, E. Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[14] Q. Zhang. High-dimensional mediation analysis with applications to causal gene identification. *Statistics in Biosciences*, pages 1–20, 2021.

[15] R. R. Zhou, L. Wang, and S. D. Zhao. Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3):573–589, 2020.

[16] S. Zhou, S. van de Geer, and P. Bühlmann. Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*, 2009.

Web Table 1: Average computational time (in seconds) for the simulation settings $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$ and $(400, 5, 0, 0.5)$

(q, p_m, r, τ)	$(5, 1, 0, 0.5)$						
Method	MedDiC	GLLZ	HIMA	HIMA2	ZWZ	MedR L1	MedR L2
Average time	2.8	1.1	68.1	159.5	365.4	2.5×10^3	1.6×10^4

(q, p_m, r, τ)	$(400, 5, 0, 0.5)$		
Method	MedDiC	GLLZ AMS	GLLZ US
Average time	29.6	1.5	471.4

Web Table 2: Results of Maize data: Number of SNPs selected with each type of effect and from each analysis at $FDR = 0.2$.

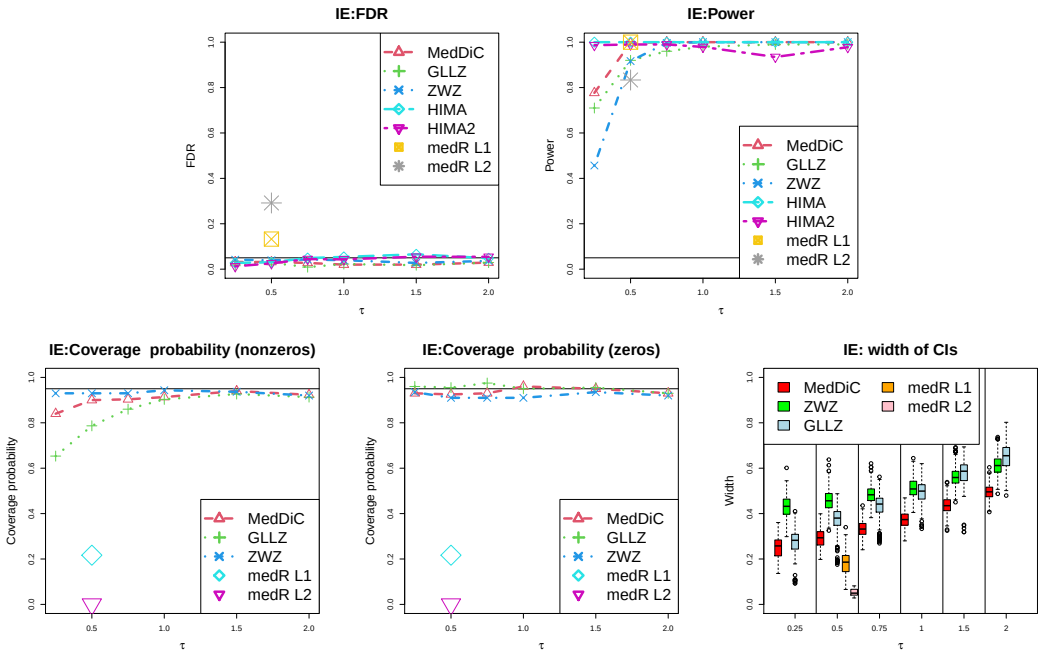
	DS	DT	GDS	GDT
Indirect Effect (MedDiC)	233	551	50	48
Direct Effect (MedDiC)	11	1	31	50
Total Effect (MedDiC)	25	62	37	46
Indirect Effect (GLLZ-US)	0	0	618	721
Direct Effect (GLLZ-US)	1915	2175	2266	1953
Indirect Effect (MedR L1)	0	0	0	0
Direct Effect (MedR L1)	0	0	0	0
Total Effect (MedR L1)	0	0	0	0

Web Table 3: Results of Maize data: Number of iSNPs (top 100) with HiChIP connections to the candidate mediator genes.

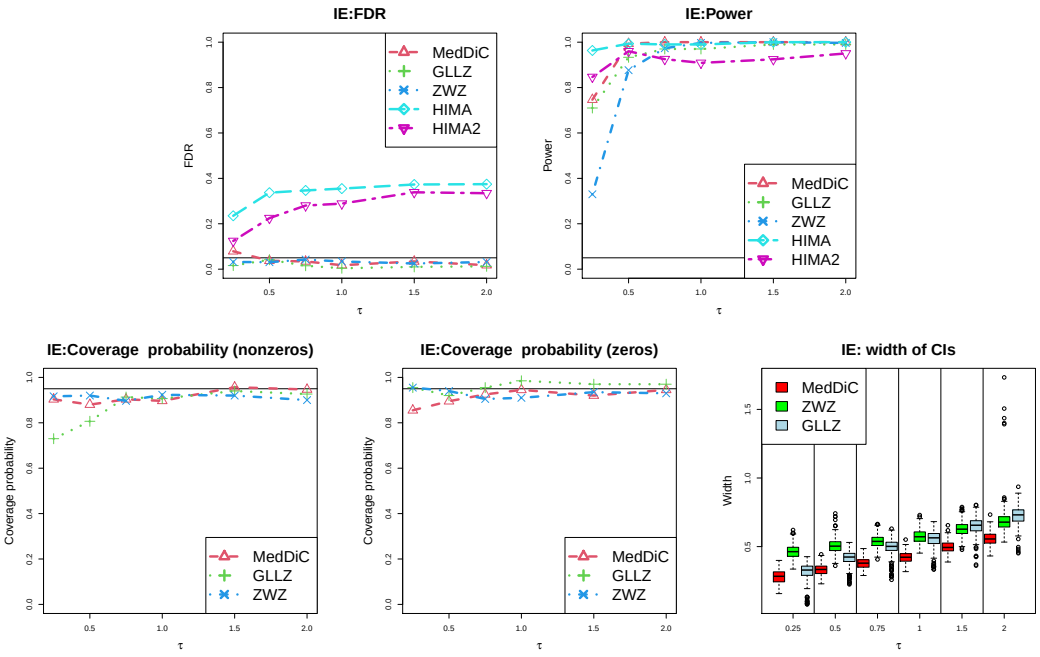
	DS	DT	GDS	GDT
Indirect Effect (MedDiC)	6	4	6	6
Indirect Effect (GLLZ US)	3	3	3	3
Indirect Effect (MedR L1)	9	4	9	7

Web Table 4: Results of Mouse data: Number of cis-regulated genes selected with each type of effect from each analysis, and their overlaps across phenotypes.

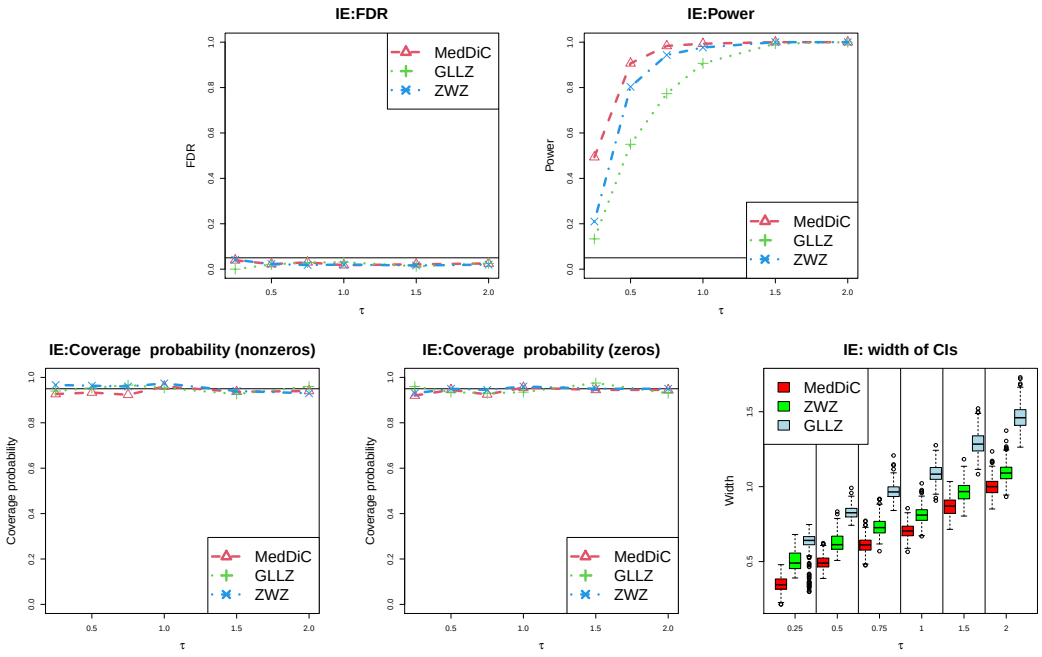
	10wk	10wkRep	rbm	10wk, 10wkRep	10wk, rbm	10wkRep, rbm
Indirect Effect (MedDiC)	40	46	44	30	21	24
Direct Effect (MedDiC)	2	3	5	2	1	1
Total Effect(MedDiC)	22	14	24	12	15	10
Indirect Effect (ZWZ)	13	31	16	12	10	14
Direct Effect (ZWZ)	24	13	15	12	14	11
Indirect Effect (GLLZ)	0	0	0	0	0	0
Direct Effect (GLLZ)	1	1	2	1	1	1
Indirect Effect (MedR L1)	20	15	20	14	16	14
Direct Effect (MedR L1)	0	0	0	0	0	0
Total Effect (MedR L1)	20	15	20	14	16	14



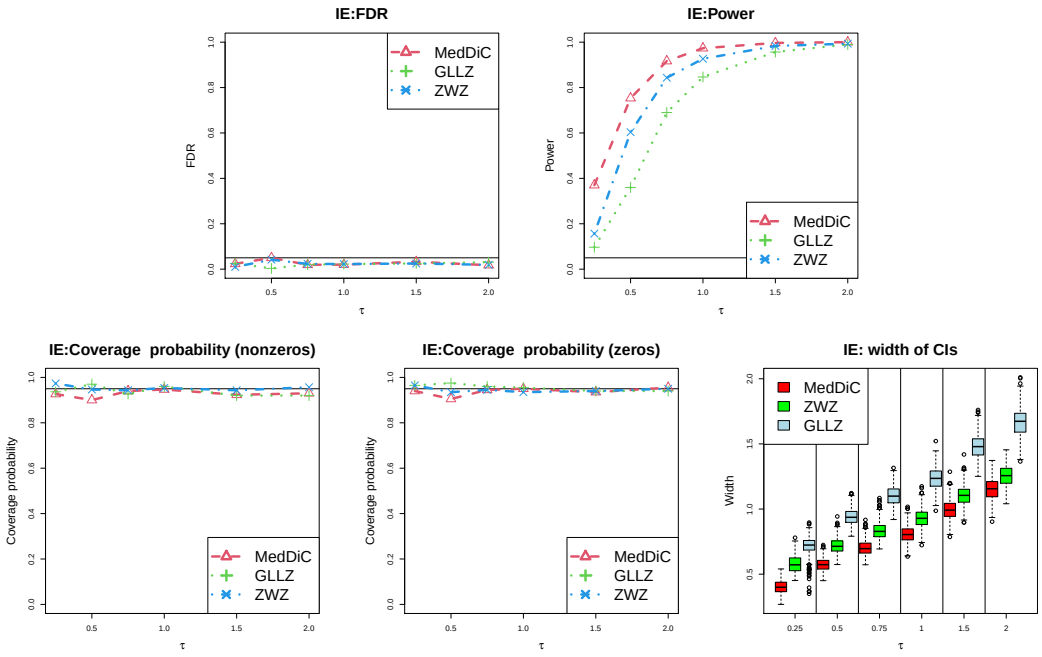
Web Figure 1: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



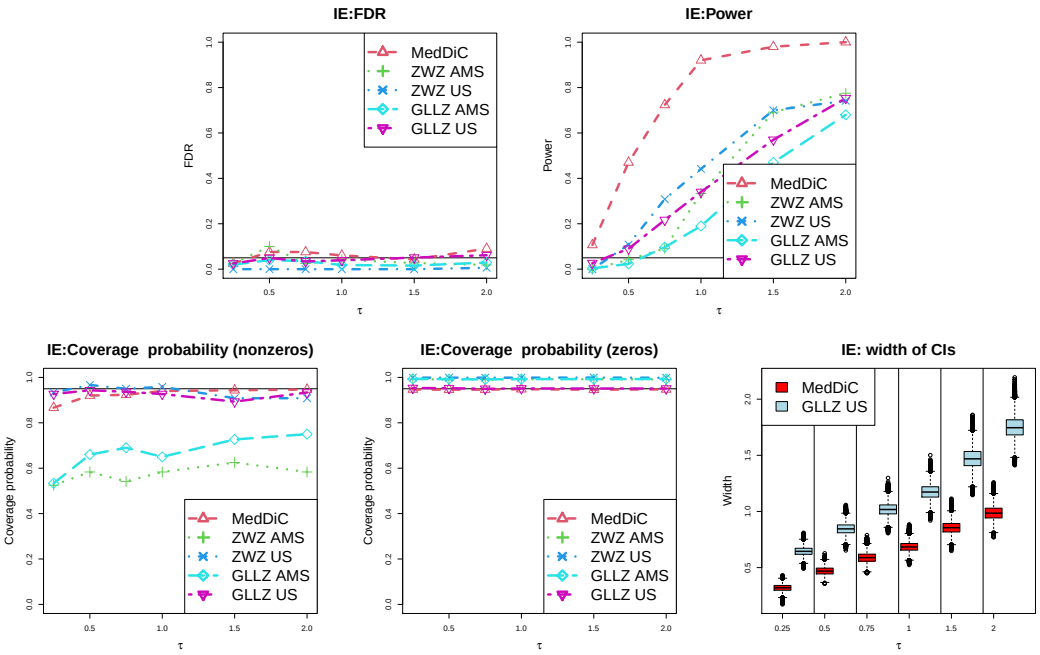
Web Figure 2: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



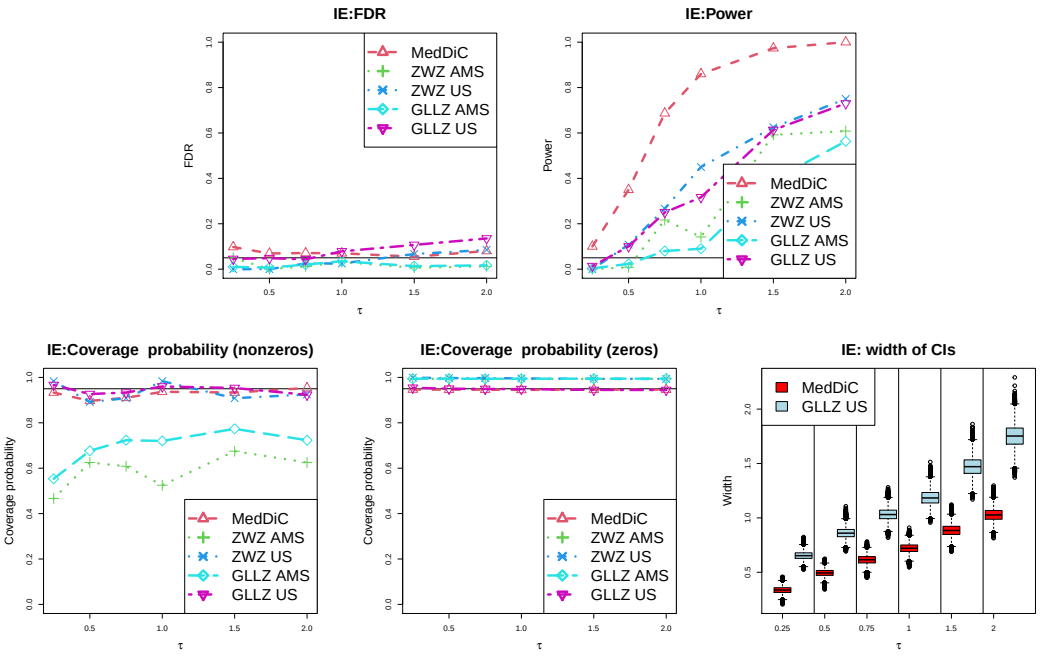
Web Figure 3: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



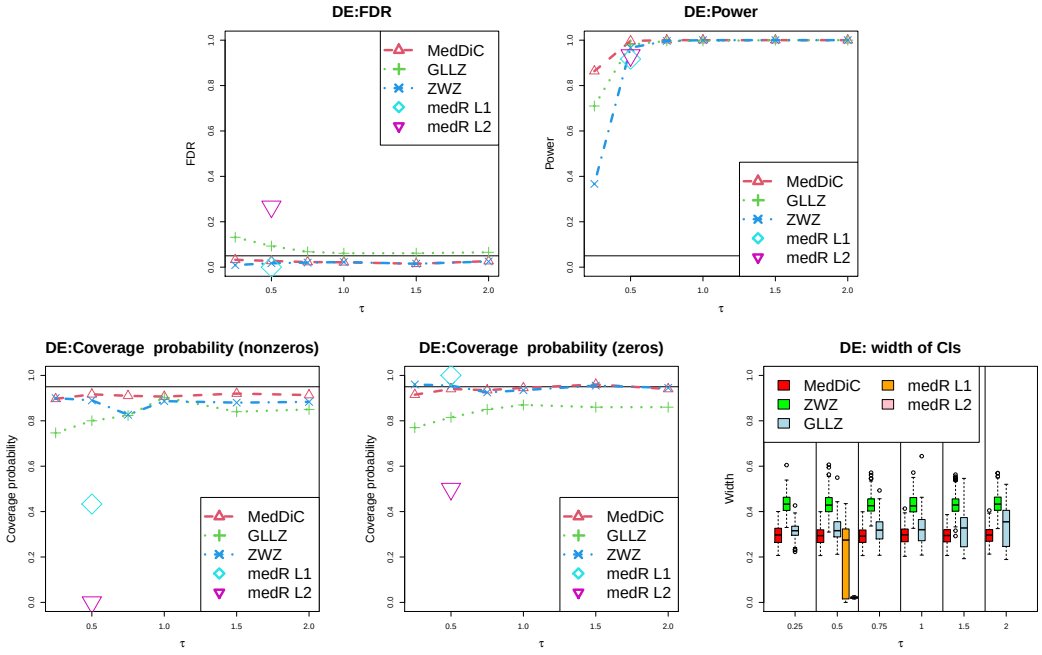
Web Figure 4: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



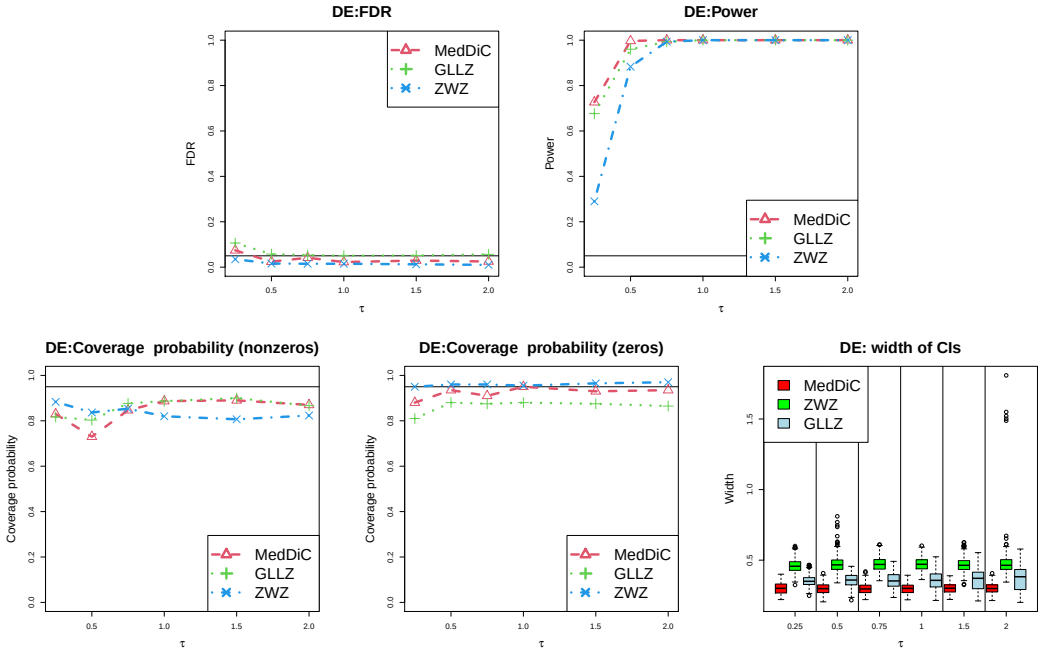
Web Figure 5: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



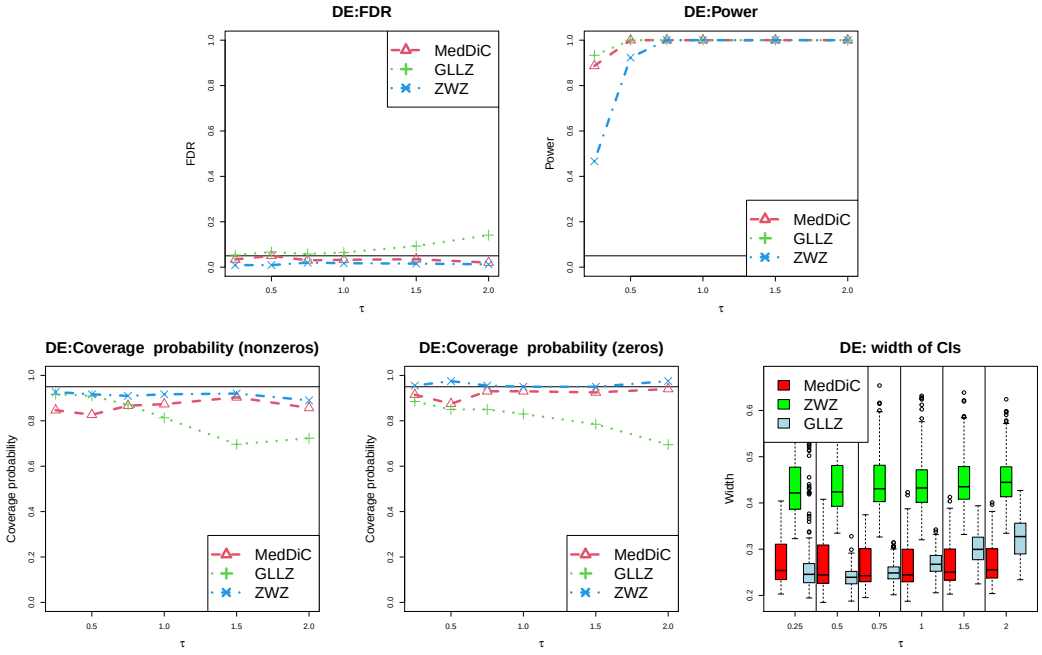
Web Figure 6: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



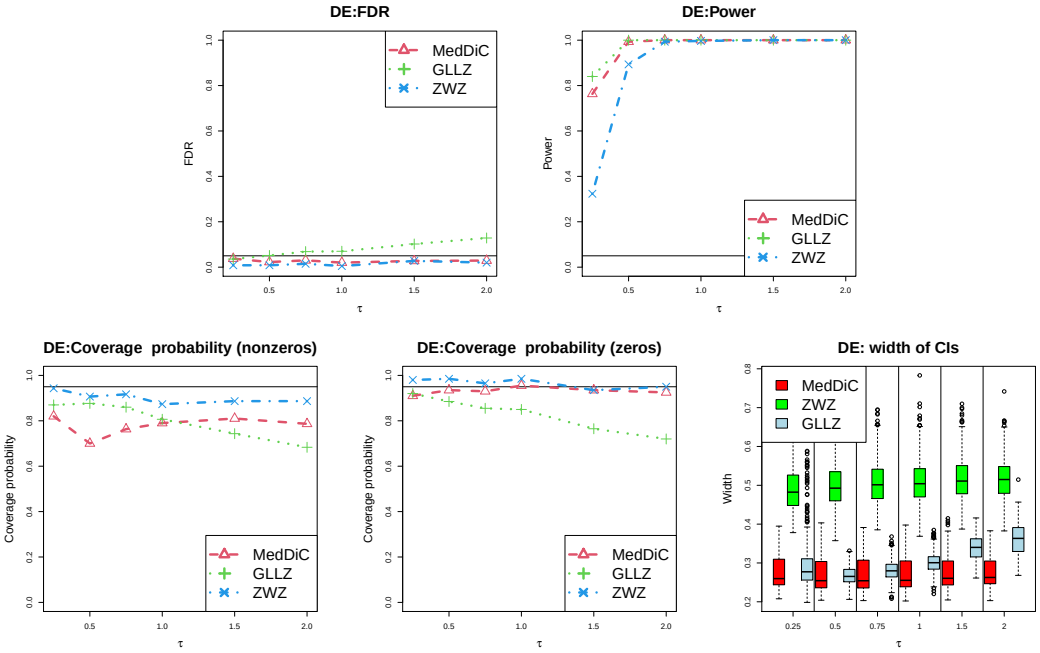
Web Figure 7: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



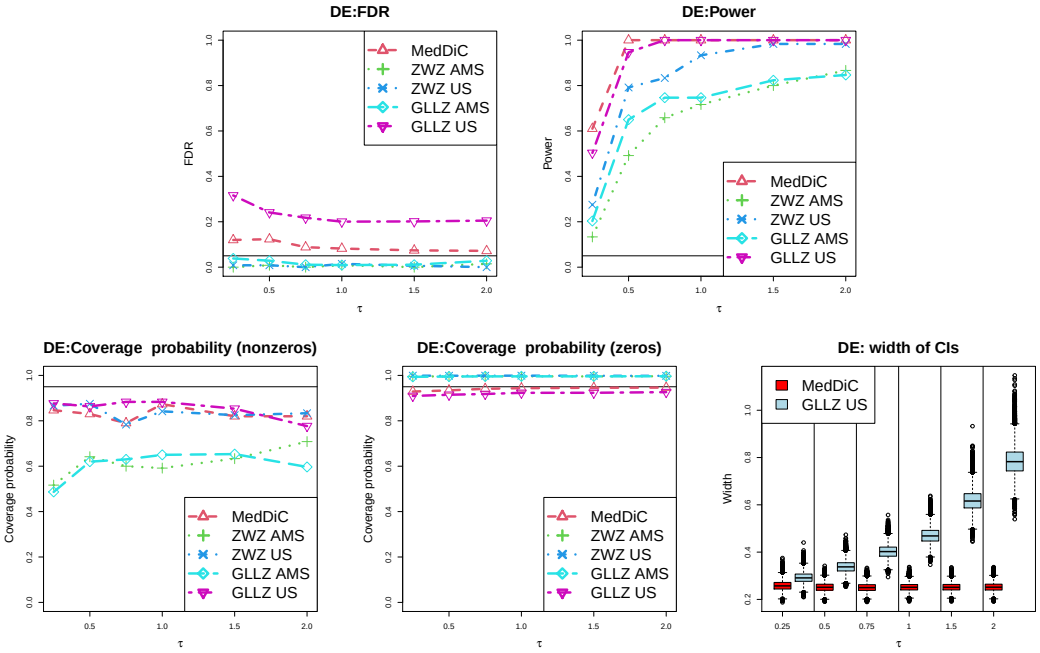
Web Figure 8: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



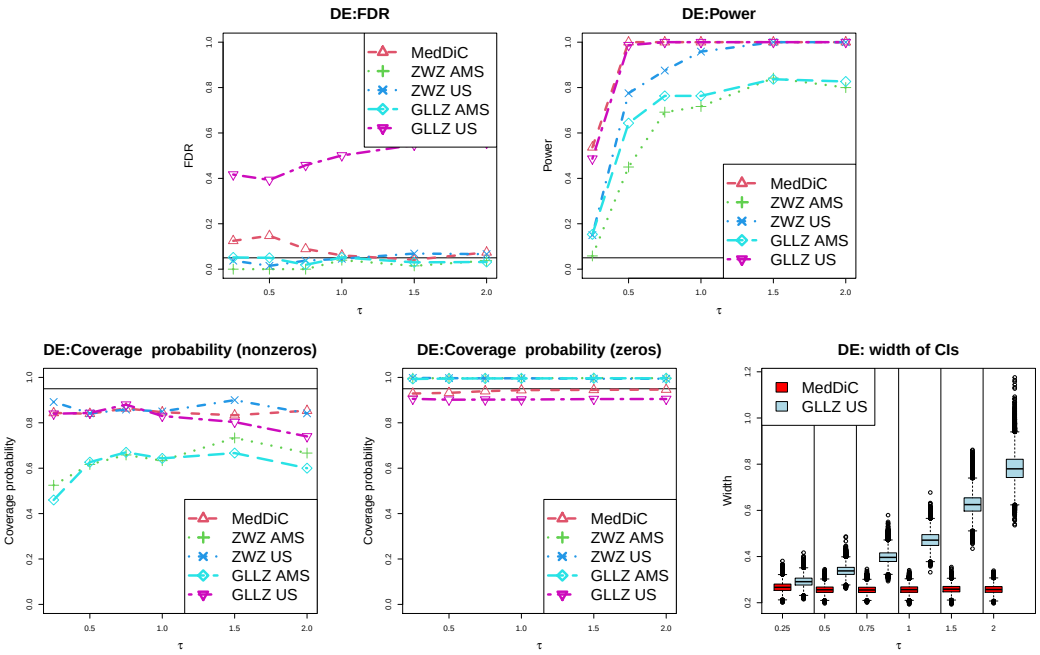
Web Figure 9: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



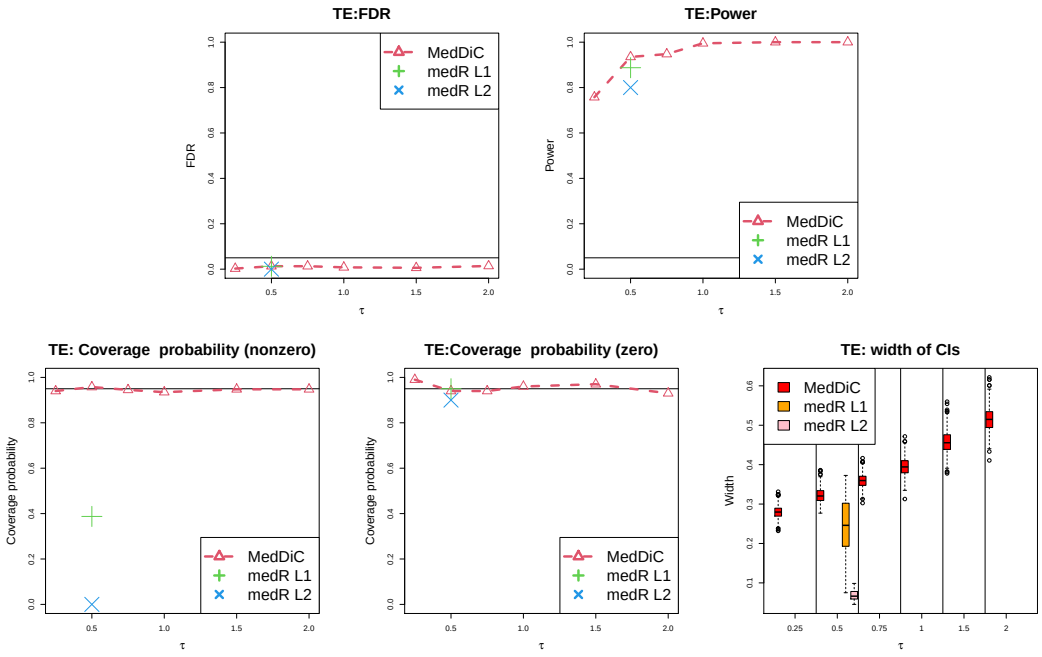
Web Figure 10: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



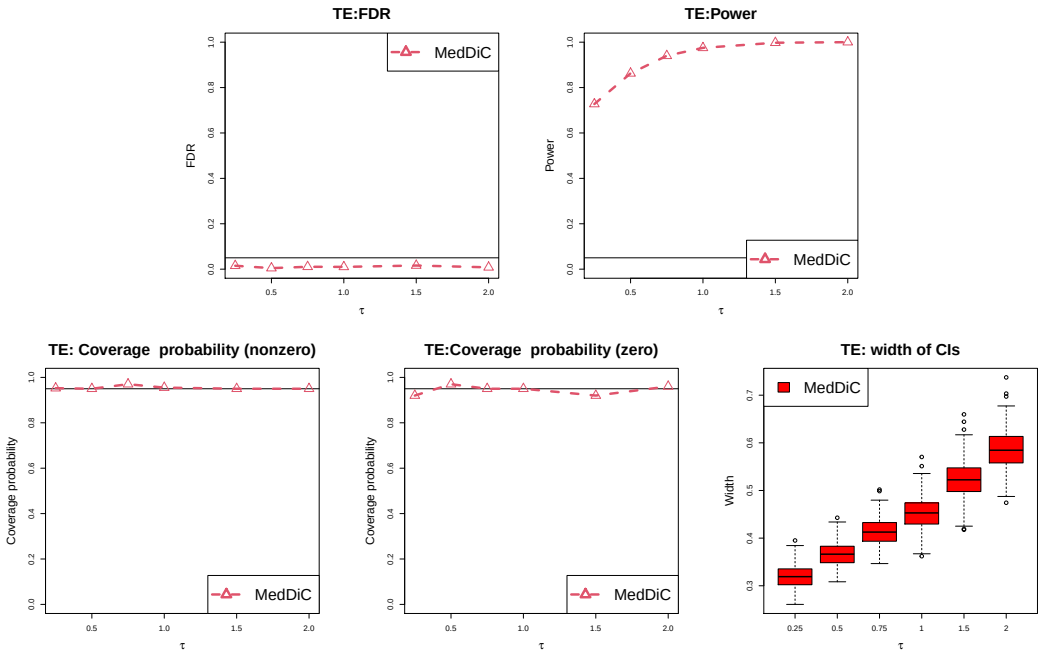
Web Figure 11: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



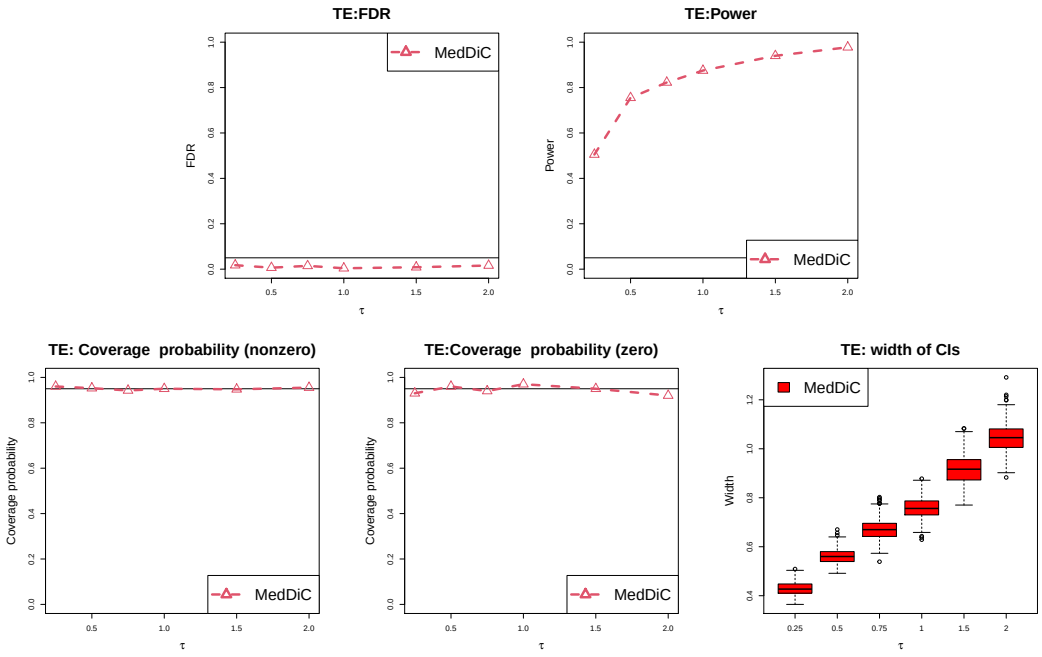
Web Figure 12: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



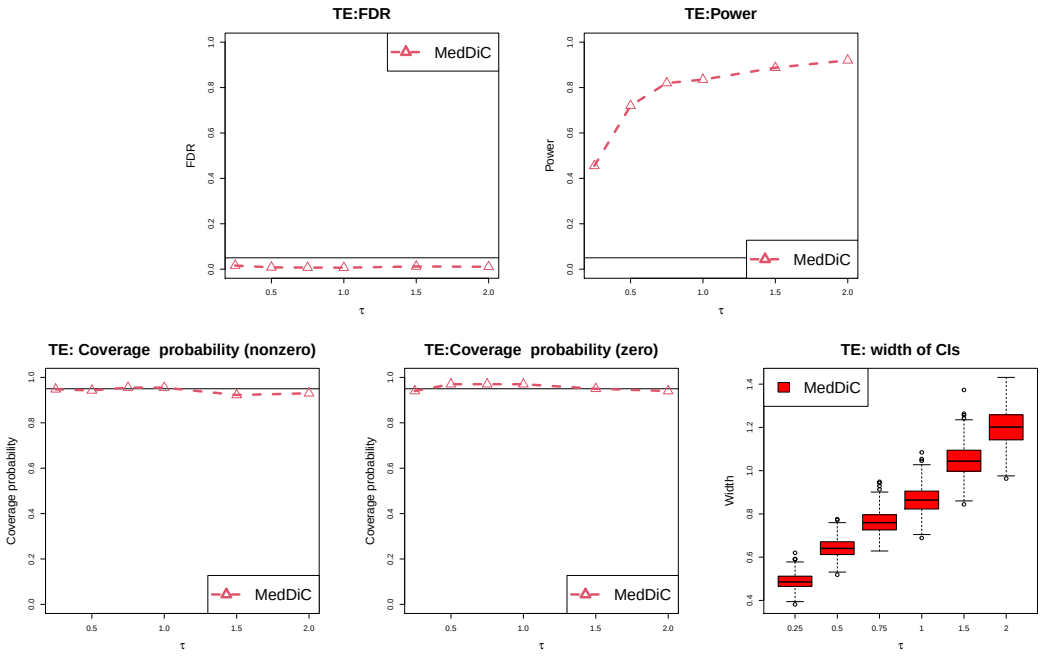
Web Figure 13: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



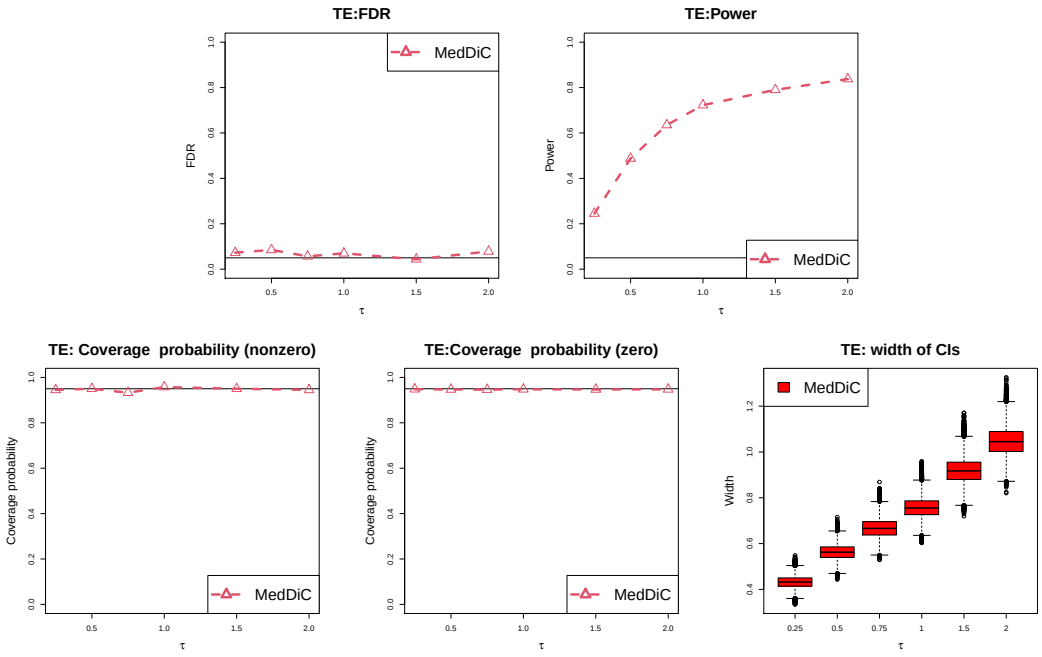
Web Figure 14: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



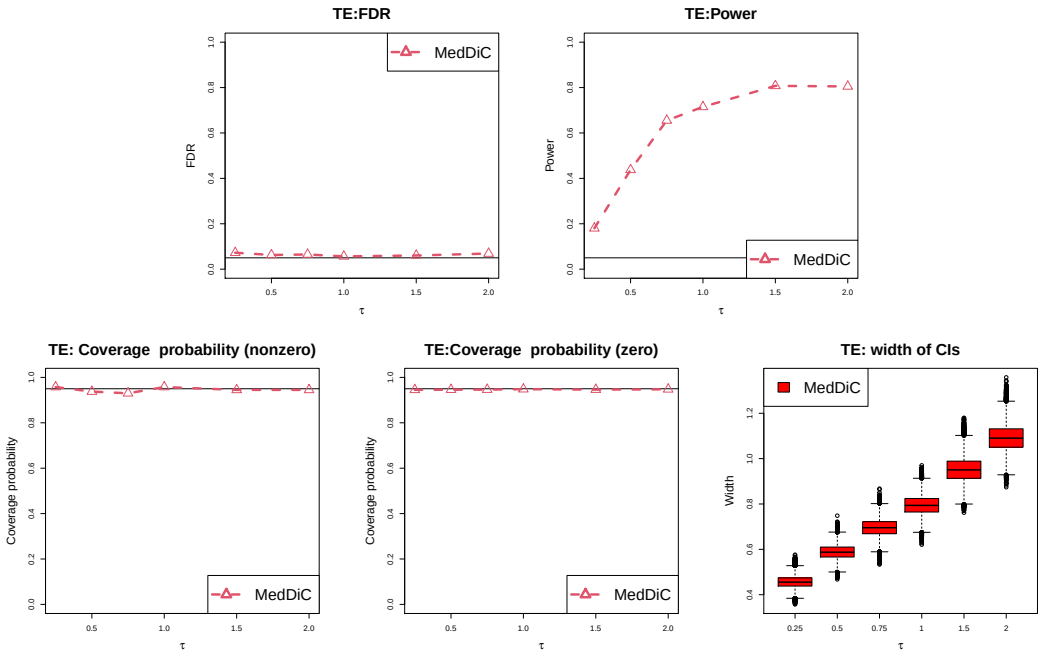
Web Figure 15: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



Web Figure 16: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



Web Figure 17: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



Web Figure 18: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).