Supporting Information for High Dimensional Mediation Analysis with Latent Variables

Andriy Derkach*, Ruth M. Pfeiffer, Ting Huei Chen, Joshua N. Sampson

In the supplementary information, we describe the EM algorithm for maximizing the penalized likelihoods in equation (18), prove the three main propositions, and provide other details and figures that were omitted from the main document.

Web Appendix A Identifiability of total, natural direct and natural indirect effects

The total effect (TE), natural direct effect (NDE), and natural indirect effect (IDE), as defined by equations (1-3) of the main text, are not generally identifiable [1]. However, we show that the three effects are identifiable when the models defined by equations (4-7) hold. Moreover, although not discussed in the main text, we can show that the three effects are identifiable even without the assumption of conditional independence (i.e $\boldsymbol{e}_{f,i} \sim N(0, I_q)$) among the factors. In other words, we prove that the effects are identifiable when the true underlying model is defined by equations (A.1 - A.4) below and Σ_q^* is of full rank. We assume that the distribution of Y_i belongs to an exponential family,

$$f(Y_i; \zeta_i^*, \psi_Y^*) = \exp\left[\left\{Y_i\zeta_i^* - b(\zeta_i^*)\right\}/a(\psi_Y^*) + c(Y_i, \psi_Y^*)\right] \tag{A.1}$$

with

$$\zeta_i^* = \gamma_Y^* + \beta_{EY}^* E_i + \boldsymbol{\beta}_{FY}^{*'} \boldsymbol{F}_i^*. \tag{A.2}$$

We also assume that $\boldsymbol{F}_i^*$ and $\boldsymbol{M}_i$ are normally distributed:

$$\boldsymbol{F}_i^* = \boldsymbol{\beta}_{EF}^* E_i + \boldsymbol{e}_{f,i}^* \text{ with } \boldsymbol{e}_{f,i}^* \sim N(0, \Sigma_q^*), \tag{A.3}$$

where Σ_q^* is the q by q covariance matrix of full rank and

$$\boldsymbol{M}_i = \boldsymbol{\gamma}_M^* + \Lambda^* \boldsymbol{F}_i^* + \boldsymbol{e}_{m,i}^* \text{ with } \boldsymbol{e}_{m,i}^* \sim N(0, \Psi^{*2}), \Psi^{*2} = diag(\psi_1^{*2}, ..., \psi_p^{*2}). \tag{A.4}$$

Moreover, the estimates of these effects from fitting the models assuming independence (i.e. $e_{f,i} \sim N(0, I_q)$) are consistent for the true estimates regardless of the format of Σ_q^*.

Under the assumption of sequential ignorability, we know that the key terms in equations (1-3) defining the TE, NDE, and NIE are related to the relevant probability distributions by

$$\mathbb{E}\left[Y_i\{e', \boldsymbol{F}^*(e)\}\right] = \int_{\boldsymbol{x}} \int_{\boldsymbol{f}^*} \mathbb{E}\left(Y_i | \boldsymbol{F}_i^* = \boldsymbol{f}^*, E_i = e', \boldsymbol{X}_i = \boldsymbol{x}\right) dF_{\boldsymbol{F}_i^* | E_i = e, X_i = x}(\boldsymbol{f}^*) dF_{X_i}(x), \tag{A.5}$$

where $\boldsymbol{X}_i$ is the set of baseline covariates, $dF_{\boldsymbol{F}_i^* | E_i = e, X_i = x}(\boldsymbol{f}^*)$ is the conditional distribution of latent factors $\boldsymbol{F}_i^*$ and $dF_{X_i}(x)$ is the distribution of baseline covariates. Under a generalized linear model, the conditional mean of Y is

$$\mathbb{E}\left(Y_i | \boldsymbol{F}_i^* = \boldsymbol{f}^*, E_i = e', \boldsymbol{X}_i = \boldsymbol{x}\right) = g^{-1}(\gamma_Y^* + \beta_{EY}^* e' + \boldsymbol{\beta}_{FY}^{*'} \boldsymbol{f}^* + \boldsymbol{\beta}_{XY}^{*'} \boldsymbol{x}).,\tag{A.6}$$

where $g()$ is a link function. We now note that equation A.5 is equivalent to

$$\mathbb{E}\left[Y_i\{e', \boldsymbol{F}^*(e)\}\right] = \int_{\boldsymbol{x}} \int_{\boldsymbol{f}^{**}} g^{-1}(\gamma_Y^* + \beta_{EY}^* e' + f^{**} + \boldsymbol{\beta}_{XY}^{*'} \boldsymbol{x}) dF_{F_i^{**} | E_i = e, X_i = x}(f^{**}) dF_{X_i}(x), \tag{A.7}$$

where $f^{**} = \boldsymbol{\beta}_{FY}^{*'} f^* = \beta_{FY}^{*'} \Sigma_q^{*1/2} \Sigma_q^{*-1/2} \boldsymbol{f}^* = \boldsymbol{\beta}_{FY}^{'} \boldsymbol{f}$ and

$$\boldsymbol{F}_i = \Sigma_q^{*-1/2} \boldsymbol{F}_i^* = \Sigma_q^{*-1/2} \boldsymbol{\beta}_{EF}^* E_i + \boldsymbol{e}_{f,i} = \boldsymbol{\beta}_{EF} E_i + \boldsymbol{e}_{f,i} \text{ with } \boldsymbol{e}_{f,i}^* \sim N(0, I_q).$$

Furthermore, the parameters from the model defined by (4 - 7), $\boldsymbol{\beta}_{EF}$ and $\boldsymbol{\beta}_{FY}$ are identifiable up to orthogonal rotation and $\boldsymbol{\beta}_{FY}^{'}\boldsymbol{\beta}_{EF}$, $\boldsymbol{\beta}_{FY}^{'}\Omega_{XF}$ and $\boldsymbol{\beta}_{FY}^{'}\boldsymbol{\beta}_{FY}$ are identifiable (see Proposition 1). Therefore, the conditional distribution of $F_i^{**} = \boldsymbol{\beta}_{FY}^{'}\boldsymbol{F}_i$ is

$$F_i^{**}|E_i = e, \boldsymbol{X}_i = \boldsymbol{x} = \boldsymbol{\beta}_{FY}^{'}\boldsymbol{F}_i|E_i = e, \boldsymbol{X}_i = \boldsymbol{x} \sim N\left(\boldsymbol{\beta}_{FY}^{'}\boldsymbol{\beta}_{EF}e + \boldsymbol{\beta}_{FY}^{'}\Omega_{XF}\boldsymbol{x}, \boldsymbol{\beta}_{FY}^{'}\boldsymbol{\beta}_{FY}\right).$$

Thus, expectation A.5 is identifiable.

Web Appendix B Derivation of the likelihoods

B.1 Prospective likelihood

Based on the model in (4-7), the conditional distribution of $Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i$ in (9) is multivariate normal with mean vector $(\mu_Y, \boldsymbol{\mu}_M, \boldsymbol{\mu}_F)$, where

$$\mu_Y = \gamma_Y + (\boldsymbol{\beta}'_{FY}\boldsymbol{\beta}_{EF} + \beta_{EY})E_i,\ \boldsymbol{\mu}_M = \boldsymbol{\gamma}_M + \Lambda\boldsymbol{\beta}_{EF}E_i,\ \boldsymbol{\mu}_F = \boldsymbol{\beta}_{EF}E_i, \tag{B.1}$$

and covariance matrix

$$\Sigma_{Y,M,F} = \begin{bmatrix} \boldsymbol{\beta}'_{FY}\boldsymbol{\beta}_{FY} + \sigma_Y^2 & \boldsymbol{\beta}'_{FY}\Lambda' & \boldsymbol{\beta}'_{FY} \\ \Lambda\boldsymbol{\beta}_{FY} & \Lambda\Lambda' + \Psi^2 & \Lambda \\ \boldsymbol{\beta}_{FY} & \Lambda' & I \end{bmatrix}. \tag{B.2}$$

We also use following definitions of the matrices $\Sigma_{F(Y,M)}$ and $\Sigma_{Y,M}$,

$$\Sigma_{F(Y,M)} = Cov(F, (Y, M)|E) = \begin{bmatrix} \boldsymbol{\beta}_{FY} & \Lambda' \end{bmatrix}, \tag{B.3}$$

$$\Sigma_{Y,M} = Cov(Y, M|E) = \begin{bmatrix} \boldsymbol{\beta}'_{FY}\boldsymbol{\beta}_{FY} + \sigma_Y^2 & \boldsymbol{\beta}'_{FY}\Lambda' \\ \Lambda\boldsymbol{\beta}_{FY} & \Lambda\Lambda' + \Psi^2 \end{bmatrix}. \tag{B.4}$$

B.2 Retrospective likelihood

Here, we assume $E_i \sim N(\gamma_E, \sigma_E^2)$. Then under the rare disease assumption, the joint distribution of $\boldsymbol{M}_i, \boldsymbol{F}_i$ and E_i in controls is multivariate normal with mean vector $(\gamma_E, \boldsymbol{\mu}_M, \boldsymbol{\mu}_F)$,

$$\boldsymbol{\mu}_M = \boldsymbol{\gamma}_M + \Lambda \boldsymbol{\beta}_{EF} \gamma_E,\ \boldsymbol{\mu}_F = \boldsymbol{\beta}_{EF} \gamma_E,\ \mu_E = \gamma_E, \tag{B.5}$$

and covariance matrix

$$\Sigma_{E,M,F} = \begin{bmatrix} \sigma_E^2 & \sigma_E^2 \boldsymbol{\beta}'_{EF} \Lambda' & \sigma_E^2 \boldsymbol{\beta}'_{EF} \\ \sigma_E^2 \Lambda \boldsymbol{\beta}_{EF} & \sigma_E^2 \Lambda \boldsymbol{\beta}_{EF} \boldsymbol{\beta}'_{EF} \Lambda' + \Lambda\Lambda' + \Psi^2 & \sigma_E^2 \Lambda \boldsymbol{\beta}_{EF} \boldsymbol{\beta}'_{EF} + \Lambda \\ \sigma_E^2 \boldsymbol{\beta}_{EF} & \sigma_E^2 \boldsymbol{\beta}_{EF} \boldsymbol{\beta}'_{EF} \Lambda' + \Lambda' & \sigma_E^2 \boldsymbol{\beta}_{EF} \boldsymbol{\beta}'_{EF} + I \end{bmatrix}. \tag{B.6}$$

We further define the matrices $\Sigma_{F(E,M)}$ and $\Sigma_{E,M}$

$$\Sigma_{E(M,F)} = Cov(F, (E, M)) = \begin{bmatrix} \sigma_E^2 \boldsymbol{\beta}_{EF} & \sigma_E^2 \boldsymbol{\beta}_{EF} \boldsymbol{\beta}'_{EF} \Lambda' + \Lambda' \end{bmatrix}, \text{ and} \tag{B.7}$$

$$\Sigma_{E,M} = Cov(E, M) = \begin{bmatrix} \sigma_E^2 & \sigma_E^2 \boldsymbol{\beta}'_{EF} \Lambda' \\ \sigma_E^2 \Lambda \boldsymbol{\beta}_{EF} & \sigma_E^2 \Lambda \boldsymbol{\beta}_{EF} \boldsymbol{\beta}'_{EF} \Lambda' + \Lambda\Lambda' + \Psi^2 \end{bmatrix}. \tag{B.8}$$

Lastly, the joint distribution of $\boldsymbol{M}_i, \boldsymbol{F}_i$ and E_i in cases is also multivariate normal with the covariance matrix $\Sigma_{E,M,F}$, but a different mean vector $(\mu_E^1, \boldsymbol{\mu}_M^1, \boldsymbol{\mu}_F^1)$,

$$\begin{bmatrix} \mu_E^1 \\ \boldsymbol{\mu}_M^1 \\ \boldsymbol{\mu}_F^1 \end{bmatrix} = \begin{bmatrix} \gamma_E \\ \boldsymbol{\gamma}_M + \Lambda \boldsymbol{\beta}_{EF} \gamma_E \\ \boldsymbol{\beta}_{EF} \gamma_E \end{bmatrix} + \Sigma_{EMF} \begin{bmatrix} \beta_{EY} \\ \mathbf{0} \\ \boldsymbol{\beta}_{FY} \end{bmatrix}. \tag{B.9}$$

Web Appendix C EM algorithms to obtain $\hat{\boldsymbol{\theta}}_P^*$

C.1 Basic EM steps

We use a fast coordinate descent algorithm in the maximization step of the EM algorithm [8, 10] based on the approaches developed for penalized factor analysis [5, 6, 12]. First, we outline of the EM algorithm and then discuss the details separately for prospective and retrospective sampling.

Recall, for a fixed set of regularization parameters, $\{\rho_1, \rho_2, \rho_3\}$, the goal of the $(k+1)^{th}$ iteration of the M-step is to find the parameters that maximize the expected value of the penalized full likelihood (EPFLL), where the expectation is computed using the parameters from the k^{th} iteration:

$$EPFLL(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^k) = \mathbb{E}\left(L^F(\boldsymbol{\theta})|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k\right) - \rho_1 \sum_{j=1}^{q} P(\beta_{FY,j}) - \rho_2 \sum_{j=1}^{q} P(\beta_{EF,j}) - \rho_3 \sum_{m=1}^{p}\sum_{j=1}^{q} P(\lambda_{mj}). \quad \text{(C.1)}$$

Under prospective sampling, the full likelihood can be decomposed into three independent parts (see 10) that are maximized separately at each M-step

$$\underset{\gamma_Y, \boldsymbol{\beta}_{FY}, \beta_{EY}, \psi_y}{\arg\max} \left(\sum_{i=1}^{N} \mathbb{E}\left[\log\left\{f_Y\left(Y_i|E_i, \boldsymbol{M}_i, \boldsymbol{F}_i\right)\right\} \;|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k\right] - \rho_1 \sum_{j=1}^{q} \frac{|\beta_{FY,j}|}{|\hat{\beta}^0_{FY,j}|} \right), \quad \text{(C.2)}$$

$$\underset{\boldsymbol{\beta}_{EF}}{\arg\max} \left(\sum_{i=1}^{N} \mathbb{E}\left[\log\left\{f_F\left(\boldsymbol{F}_i|E_i\right)\right\} |Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k\right] - \rho_2 \sum_{j=1}^{q} \frac{|\beta_{EF,j}|}{|\hat{\beta^0}_{EF,j}|} \right), \quad \text{(C.3)}$$

and

$$\underset{\boldsymbol{\gamma}_M, \Lambda, \Psi^2}{\arg\max} \left(\sum_{i=1}^{N} \mathbb{E}\left[\log\left\{f_M\left(\boldsymbol{M}_i|\boldsymbol{F}_i\right)\right\} |Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k\right] - \rho_3 \sum_{m=1}^{p}\sum_{j=1}^{q} \frac{|\lambda_{mj}|}{|\hat{\lambda}^0_{mj}|} \right). \quad \text{(C.4)}$$

We note that each objective function (C.2 - C.4) is non-convex in $\gamma_Y, \boldsymbol{\beta}_{FY}, \beta_{EY}, \psi_Y, \boldsymbol{\beta}_{EF}, \boldsymbol{\gamma}_M, \Lambda, \Psi^2, \gamma_E$ and σ_E^2. However, each objective function is convex when the variances ψ_Y, σ_E^2 and Ψ^2 are fixed. Therefore, we fix them at the previous estimates $\hat{\psi}_Y^k$, $\hat{\sigma}_E^{2k}$ and $\hat{\Psi}^{2k}$ and solve the three convex optimization problems sequentially. After the regression coefficients are updated, the corresponding variance/dispersion

parameters are updated.

For retrospective sampling, we modify these equations slightly (see Web Appendix C.3). Parameters $\boldsymbol{\gamma}_M$, Λ and Ψ^2 are estimated from the same equation $C.4$. The remaining parameters can be updated by maximizing the joint function of E_i and $\boldsymbol{M}_i$,

$$\underset{\gamma_E,\beta_{EY},\sigma_E^2,\boldsymbol{\beta}_{FY},\boldsymbol{\beta}_{EF}}{\arg\max} \left(\sum_{i=1}^{N} \mathbb{E}\left[\log\left\{ f_{EF}\left(E_i, \boldsymbol{F}_i\right)\right\} | Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k \right] - \rho_1 \sum_{j=1}^{q} \frac{|\beta_{FY,j}|}{|\hat{\beta}^0_{FY,j}|} - \rho_2 \sum_{j=1}^{q} \frac{|\beta_{EF,j}|}{|\hat{\beta}^0_{EF,j}|} \right). \quad \text{(C.5)}$$

The objective function (C.5) can not be decomposed into two separate optimization problems because the mean is a product of the parameters (see B.9). Additionally it is non convex in $\boldsymbol{\beta}_{FY}, \beta_{EY}, \boldsymbol{\beta}_{EF}, \gamma_E$ and σ_E^2. However, this objective function is convex when $\boldsymbol{\beta}_{EF}$ or $\boldsymbol{\beta}_{FY}$ is fixed. We therefore update $\boldsymbol{\beta}_{EF}$ from (C.5) while keeping all other parameters fixed and update the remaining parameters while $\boldsymbol{\beta}_{EF}$ is kept fixed.

At the $(k+1)^{th}$ E-step, we calculate two conditional expectations of the latent variable F_i given Y_i, $\boldsymbol{M}_i$ and E_i (see [5, 6, 10, 12]) used in the following M step:

$$\boldsymbol{E}_{f_i} = \mathbb{E}\left(\boldsymbol{F}_i | \boldsymbol{M}_i, E_i, Y_i; \hat{\boldsymbol{\theta}}^k \right) \quad \text{(C.6)}$$

and

$$\boldsymbol{E}_{f_i^2} = \mathbb{E}(\boldsymbol{F}_i \boldsymbol{F}_i' | Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k) = Var(\boldsymbol{F}_i | Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k) + \boldsymbol{E}_{f_i} \boldsymbol{E}_{f_i}'. \quad \text{(C.7)}$$

We assume that either the distribution of $(\boldsymbol{F}_i, Y_i, \boldsymbol{M}_i)$ given E_i, or $(\boldsymbol{F}_i, E_i, \boldsymbol{M}_i)$ given Y_i is multivariate normal to obtain closed form solutions. For other joint distributions numerical integration is required for the for the E-step [11]. As a result, under prospective sampling, the conditional expectations of the latent variables are

$$
\begin{aligned}
\boldsymbol{E}_{f_i} &= \hat{\boldsymbol{\beta}}^k_{EF} E_i + \begin{bmatrix} \hat{\boldsymbol{\beta}}^k_{FY} & \hat{\Lambda}^{k'} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}^{k'}_{FY}\hat{\boldsymbol{\beta}}^k_{FY} + \hat{\sigma}^{2k}_Y & \hat{\sigma}^{2k}_E \hat{\boldsymbol{\beta}}^{k'}_{FY}\hat{\Lambda}^{k'} \\ \hat{\Lambda}^k\hat{\boldsymbol{\beta}}^k_{FY} & \hat{\Lambda}^k\hat{\Lambda}^{k'} + \hat{\Psi}^{2k} \end{bmatrix}^{-1} \begin{bmatrix} Y_i - \hat{\gamma}^k_Y - (\hat{\boldsymbol{\beta}}^{k'}_{FY}\hat{\boldsymbol{\beta}}^k_{EF} + \hat{\beta}^k_{EY})E_i \\ \boldsymbol{M}_i - \hat{\boldsymbol{\gamma}}^k_M - \hat{\Lambda}^{k'}\hat{\boldsymbol{\beta}}^k_{EF}E_i \end{bmatrix} \\
&= \boldsymbol{\mu}^k_F + \Sigma^k_{F(YM)}\Sigma^{-1k}_{Y,M} \begin{bmatrix} Y_i - \mu^k_Y \\ \boldsymbol{M}_i - \boldsymbol{\mu}^k_M \end{bmatrix}
\end{aligned}
$$

and

$$
\boldsymbol{E}_{f_i^2} = I - \Sigma^k_{F(YM)}\Sigma^{-1k}_{Y,M}\Sigma^k_{(YM)F} + \boldsymbol{E}_{f_i}\boldsymbol{E}'_{f_i},
$$

with $\Sigma^{k'}_{(YM)F} = \Sigma^k_{F(YM)}$.

Under retrospective sampling, conditional expectations of latent variables are slightly different. In controls, these expectations are

$$
\begin{aligned}
\boldsymbol{E}^0_{f_i} &= \hat{\boldsymbol{\mu}}^{0k}_F + \begin{bmatrix} \hat{\sigma}^{2k}_E\hat{\boldsymbol{\beta}}^k_{EF} & \hat{\sigma}^{2k}_E\hat{\boldsymbol{\beta}}^k_{EF}\hat{\boldsymbol{\beta}}^{k'}_{EF}\hat{\Lambda}^{k'} + \hat{\Lambda}^{k'} \end{bmatrix} \begin{bmatrix} \hat{\sigma}^{k2}_E & \hat{\sigma}^{k2}_E\hat{\boldsymbol{\beta}}^{k'}_{EF}\hat{\Lambda}^{k'} \\ \hat{\sigma}^{k2}_E\hat{\Lambda}^k\hat{\boldsymbol{\beta}}^k_{EF} & \hat{\sigma}^{k2}_E\hat{\Lambda}^k\hat{\boldsymbol{\beta}}^k_{EF}\hat{\boldsymbol{\beta}}^{k'}_{EF}\hat{\Lambda}^{k'} + \hat{\Lambda}^k\hat{\Lambda}^{k'} + \hat{\Psi}^{2k} \end{bmatrix}^{-1} \begin{bmatrix} E_i - \hat{\mu}^{0k}_E \\ \boldsymbol{M}_i - \hat{\boldsymbol{\mu}}^{0k}_M \end{bmatrix} \\
&= \hat{\boldsymbol{\mu}}^{0k}_F + \Sigma^k_{F(EM)}\Sigma^{-1k}_{E,M} \begin{bmatrix} E_i - \hat{\mu}^{0k}_E \\ \boldsymbol{M}_i - \hat{\boldsymbol{\mu}}^{0k}_M \end{bmatrix}
\end{aligned}
$$

and

$$
\boldsymbol{E}^0_{f_i^2} = I + \hat{\sigma}^{2k}_E\hat{\boldsymbol{\beta}}^k_{EF}\hat{\boldsymbol{\beta}}^{k'}_{EF} - \Sigma^k_{F(EM)}\Sigma^{-1k}_{E,M}\Sigma^k_{(EM)F} + \boldsymbol{E}^0_{f_i}\boldsymbol{E}^{0'}_{f_i},
$$

with

$$
\begin{bmatrix} \hat{\mu}^{0k}_E \\ \hat{\boldsymbol{\mu}}^{0k}_M \\ \hat{\boldsymbol{\mu}}^{0k}_F \end{bmatrix} = \begin{bmatrix} \hat{\gamma}^k_E \\ \hat{\boldsymbol{\gamma}}^k_M + \hat{\Lambda}^k\hat{\boldsymbol{\beta}}^k_{EF}\hat{\gamma}^k_E \\ \hat{\boldsymbol{\beta}}^k_{EF}\hat{\gamma}^k_E \end{bmatrix}.
$$

Lastly, in case samples, these expectations are

$$\boldsymbol{E}^1_{f_i} \;=\; \hat{\boldsymbol{\mu}}^{1k}_F + \Sigma^k_{F(EM)}\Sigma^{-1k}_{E,M}\begin{bmatrix} E_i - \hat{\mu}^{1k}_E \\ \boldsymbol{M}_i - \hat{\boldsymbol{\mu}}^{1k}_M \end{bmatrix}$$

and

$$\boldsymbol{E}^1_{f_i^2} = I + \hat{\sigma}^{2k}_E \hat{\boldsymbol{\beta}}^k_{EF}\hat{\boldsymbol{\beta}}^{k'}_{EF} - \Sigma^k_{F(EM)}\Sigma^{-1k}_{E,M}\Sigma^k_{(EM)F} + \boldsymbol{E}^1_{f_i}\boldsymbol{E}^{1'}_{f_i},$$

with

$$\begin{bmatrix} \hat{\mu}^{1k}_E \\ \hat{\boldsymbol{\mu}}^{1k}_M \\ \hat{\boldsymbol{\mu}}^{1k}_F \end{bmatrix} = \begin{bmatrix} \hat{\gamma}^k_E \\ \hat{\boldsymbol{\gamma}}^k_M + \hat{\Lambda}^k\hat{\boldsymbol{\beta}}^k_{EF}\hat{\gamma}^k_E \\ \hat{\boldsymbol{\beta}}^k_{EF}\hat{\gamma}^k_E \end{bmatrix} + \Sigma^k_{E,M,F}\begin{bmatrix} \hat{\beta}^k_{EY} \\ \mathbf{0} \\ \hat{\boldsymbol{\beta}}^k_{FY} \end{bmatrix}.$$

C.2 Further computational details for prospective sampling

We first assume that Y_i has a normal distribution and $\boldsymbol{E}_{f_i}$ and $\boldsymbol{E}_{f_i^2}$ are calculated at the E-step of the $(k+1)^{th}$ iteration. At the M-step, we then update the estimates of each set of regression coefficients (e.g. $\{\gamma_Y, \beta_{FY}, \beta_{EY}\}$, β_{FE}, $\{\Lambda, \gamma_M\}$) in sequence by the one-step coordinate descent algorithm [14]. We repeat this step until convergence of all parameters is met.

Here, we show how to update our estimate of the first set, involved in equation (C.2), and note that all other updates proceed similarly. We calculate $\hat{\gamma_Y}^{k+1}$, $\hat{\beta}^{k+1}_{EY}$, and $\hat{\boldsymbol{\beta}}^{k+1}_{FY,j}$ from the following set of sequential iterations

$$\hat{\gamma}^{k+1}_Y = \frac{\sum_{i=1}^N \left(Y_i - \hat{\beta}^k_{EY}E_i - \boldsymbol{E}'_{f_i}\hat{\boldsymbol{\beta}}^k_{FY}\right)}{N}, \tag{C.8}$$

$$\hat{\beta}^{k+1}_{EY} = \frac{\sum_{i=1}^N \left(Y_iE_i - \hat{\gamma}^{k+1}_Y E_i - E_i\boldsymbol{E}'_{f_i}\hat{\boldsymbol{\beta}}^k_{FY}\right)}{\boldsymbol{E}'\boldsymbol{E}}, \tag{C.9}$$

and

$$\hat{\beta}^{k+1}_{FY,j} = sign(\tilde{\beta}_{FY,j})\left(|\tilde{\beta}_{FY,j}| - \frac{\rho_1}{N|\hat{\beta^0_{FY,j}}|}\right)_+ \text{ for j=1,...,q,} \tag{C.10}$$

where $(\cdot)_+$ is the soft thresholding operator defined as

$$(x)_+ = \begin{cases} x, \text{ if } x > 0 \\ 0, \text{ if } x \leq 0, \end{cases} \tag{C.11}$$

and

$$\tilde{\beta}_{FY,j} = \frac{\sum_{i=1}^{N} \left(Y_i \boldsymbol{E}_{f_i}[j] - \hat{\gamma}_y^{k+1} \boldsymbol{E}_{f_i}[j] - \boldsymbol{E}_{f_i^2}[j,-j] \hat{\boldsymbol{\beta}}_{FY,-j}^{k+1} - \hat{\beta}_{EY}^{k+1} E_i \boldsymbol{E}_{f_i}[j] \right)}{\sum_{i=1}^{N} \boldsymbol{E}_{f_i^2}[j,j]}, \tag{C.12}$$

where $\boldsymbol{E}_{f_i}[j]$ denotes the j^{th} elements of the vector $\boldsymbol{E}_{f_i}$, $\boldsymbol{E}_{f_i^2}[j,j]$ is the $[j,j]$ element of the matrix and $\boldsymbol{E}_{f_i^2}[j,-j]$ is the j^{th} row without the j^{th} element of the matrix $\boldsymbol{E}_{f_i^2}$. Given these new estimates, we calculate $\hat{\sigma}_Y^{2k+1}$ by

$$\hat{\sigma}_Y^{2k+1} = \frac{\sum_{i=1}^{N} \left(Y_i - \hat{\gamma}_Y^{k+1} - \hat{\beta}_{EY}^{k+1} E_i \right)^2 - 2 \sum_{i=1}^{N} \left(Y_i - \hat{\gamma}_Y^{k+1} - \hat{\beta}_{EY}^{k+1} E_i \right) \boldsymbol{E}'_{f_i} \hat{\boldsymbol{\beta}}_{FY}^{k+1} + \hat{\boldsymbol{\beta}}_{FY}^{'k+1} \boldsymbol{E}_{f_i^2} \hat{\boldsymbol{\beta}}_{FY}^{k+1}}{N}. \tag{C.13}$$

These M and E steps are repeated until convergence is satisfied and the final estimates, $\hat{\boldsymbol{\theta}}$ are recorded.

C.3 Further computational details for retrospective sampling

The updating step for $\{\boldsymbol{\gamma}_M, \Lambda, \Psi^2\}$ is the same as for prospective sampling, with the caveat that $\boldsymbol{E}_{f_i}$ and $\boldsymbol{E}_{f_i^2}$ need to be calculated separately for cases and controls as discussed on previous section.

Here we show steps to estimate remaining parameters. In the first step we update $\boldsymbol{\beta}_{EF}$ by maximizing the penalized likelihood (C.5) with $\{\boldsymbol{\gamma}_E, \beta_{EY}, \boldsymbol{\beta}_{FY}, \sigma_E^2\}$ set to k^{th} iteration values. In the second step, we update $\{\boldsymbol{\gamma}_E, \beta_{EY}, \boldsymbol{\beta}_{FY}\}$ with $\boldsymbol{\beta}_{EF}$ set to $k+1^{th}$ iteration values (i.e. maximization of E and $\boldsymbol{M}$).

In the first step of the $(k+1)^{th}$ iteration of our coordinate descent algorithm, $\boldsymbol{\beta}_{EF}$ is updated from the conditional distribution of $\boldsymbol{F}_i$ given E_i by

$$\hat{\beta}_{EF,j}^{k+1} = sign(\tilde{\beta}_{EF,j}) \left(|\tilde{\beta}_{EF,j}| - \frac{\rho_2}{N|\hat{\beta}_{EF,j}^0|} \right)_+ \text{ for } j = 1, ..., q, \tag{C.14}$$

where $(\cdot)_+$ is the soft thresholding operator,

$$\tilde{\beta}_{EF,j} = \frac{\sum_{i=1}^{N_1}(E_i - \hat{\mu}_E^{1k})(\boldsymbol{E}_{f_i}^1[j] - \hat{\boldsymbol{\mu}}_F^{1k}[j]) + \sum_{i=1}^{N_0}(E_i - \hat{\gamma_E}^k)(\boldsymbol{E}_{f_i}^0[j] - \hat{\boldsymbol{\mu}}_F^{0k}[j])}{\sum_{i=1}^{N_1}(E_i - \hat{\mu}_E^{1k})^2 + \sum_{i=1}^{N_0}(E_i - \hat{\gamma}_E^k)^2} \text{ where} \tag{C.15}$$

$\hat{\boldsymbol{\mu}}_F^{1k}$, $\hat{\boldsymbol{\mu}}_F^{0k}$, $\hat{\mu}_E^{1k}$ and $\hat{\gamma}_E^k$ are estimates from the k^{th} iteration and $\hat{\boldsymbol{\mu}}_F^{1k}[j]$ is the j^{th} element of its vector (see section B.2). The remaining parameters γ_E, $\boldsymbol{\beta}_{FY}$, β_{EY} and σ_E^2 are estimated from the rewriting the optimization problem (C.5) as

$$\begin{aligned}\underset{\gamma_E,\boldsymbol{\beta}_{FY},\beta_{EY}}{\arg\max} \Bigg[&\sum_{i=1}^{N_1} - \mathbb{E}\left(\sum_{j=1}^{q}\frac{(F_i[j] - \hat{\boldsymbol{\beta}}_{EF}^{k+1}[j]E_i - \boldsymbol{\beta}_{FY}[j])^2}{2}|Y_i = 1, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}^k\right) \\ &- \sum_{i=1}^{N_1}\frac{(E_i - \gamma_E - \sigma_E^2\hat{\boldsymbol{\beta}}_{FE}^{k+1'}\boldsymbol{\beta}_{FY} - \sigma_E^2\beta_{EY})^2}{2\sigma_E^2} - \sum_{i=1}^{N_0}\frac{(E_i - \gamma_E)^2}{2\sigma_E^2} - \rho_1\sum_{j=1}^{q}\frac{|\beta_{FY,j}|}{|\hat{\beta}_{FY,j}^0|}\Bigg] \text{ where}\end{aligned} \tag{C.16}$$

$F[j]$, $\hat{\boldsymbol{\beta}}_{EF}^{k+1}[j]$ and $\hat{\boldsymbol{\beta}}_{FY}^{k+1}[j]$ are the j^{th} elements of the corresponding vectors. We treat (C.16) as a penalized least squares problem that is optimized by sequential iterations

$$\hat{\gamma}_E^{k+1} = \frac{\sum_{i=1}^{N_1}(E_i - \hat{\sigma}_E^{2k}\hat{\boldsymbol{\beta}}_{FE}^{k+1'}\hat{\boldsymbol{\beta}}_{FY}^k - \hat{\sigma}_E^{2k}\hat{\beta}_{EY}^k) + \sum_{i=1}^{N_0}E_i}{N_1 + N_0}, \tag{C.17}$$

$$\hat{\beta}_{EY}^{k+1} = \frac{\sum_{i=1}^{N_1}(E_i - \hat{\gamma}_E^{k+1} - \hat{\sigma}_E^{2k}\hat{\boldsymbol{\beta}}_{FE}^{k+1'}\hat{\boldsymbol{\beta}}_{FY}^k)}{\hat{\sigma}_E^{2k}N_1}, \tag{C.18}$$

and

$$\hat{\beta}_{FY,j}^{k+1} = sign(\tilde{\beta}_{FY,j})\left(|\tilde{\beta}_{FY,j}| - \frac{\rho_1}{N\hat{|\beta_{FY,j}^0|}}\right)_+ \text{ for j=1,...,q,} \tag{C.19}$$

where

$$\tilde{\beta}_{FY,j} = \frac{\sum_{i=1}^{N_1}\left(F_i[j] - \hat{\boldsymbol{\beta}}_{EF,j}^{k+1}E_i\right) + \sum_{i=1}^{N_1}\left(E_i - \hat{\gamma}_E^{k+1} - \hat{\sigma}_E^{2k}\hat{\boldsymbol{\beta}}_{FE,-j}^{k+1'}\hat{\boldsymbol{\beta}}_{FY,-j}^{k+1} - \hat{\sigma}_E^{2k}\hat{\beta}_{EY}^{k+1}\right)\hat{\boldsymbol{\beta}}_{FE,j}^{k+1}}{N_1\left(1 + \hat{\sigma}_E^{2k}(\hat{\boldsymbol{\beta}}_{FE,j}^{k+1})^2\right)}. \tag{C.20}$$

Finally, we recalculate a new value of σ_E^2 by using the updated means $(\hat{\boldsymbol{\mu}}_F^{1k+1}, \hat{\mu}_E^{1k+1}) = (\hat{\boldsymbol{\beta}}_{EF}^{k+1}\hat{\gamma}_E^{k+1}, \hat{\gamma}_E^{k+1}) + \hat{\Sigma}_{F,E}^{k+1}\hat{\boldsymbol{\beta}}_Y^{k+1}$ as

$$\hat{\sigma}_E^{2k+1} = \frac{\sum_{i=1}^{N_1}(E_i - \hat{\mu}_E^{1k+1})^2}{N_1} + \frac{\sum_{i=1}^{N_0}(E_i - \hat{\gamma}_E^{k+1})^2}{N_0}. \tag{C.21}$$

These iterations are repeated until convergence and $\hat{\boldsymbol{\theta}}^{k+1}$ is recorded as the final estimate.

Web Appendix D Proofs of Propositions 1-3

*Proof of Proposition 1 (**Identifiability**)*

Here we show that the model presented by equations (4-7) with observed data $(Y, E, \boldsymbol{M})$ is identifiable under both prospective and retrospective sampling. We show that under the condition from Theorem 5.1 of Anderson and Rubin [2] (condition 1 of proposition 1), if for all values of $Y, E, \boldsymbol{M}$, $f(\cdot;\boldsymbol{\theta}^1) = f(\cdot;\boldsymbol{\theta}^2)$ then all parameters are equal, with Λ, $\boldsymbol{\beta}_{EF}$ and $\boldsymbol{\beta}_{FY}$ identifiable up to rotation by an orthogonal matrix Q (i.e., $\Lambda^1 Q = \Lambda^2$, $Q'\boldsymbol{\beta}^1_{FY} = \boldsymbol{\beta}^2_{FY}$, $Q'\boldsymbol{\beta}^1_{EF} = \boldsymbol{\beta}^2_{EF}$). Here $f(\cdot;\boldsymbol{\theta})$ denotes the conditional density used in either in the prospective or retrospective likelihood (11 or 17). We note that the intercept γ_Y is not identifiable under retrospective sampling. In addition to the parameters $\boldsymbol{\theta}$, the mean γ_E and the variance σ^2_E of E are identifiable under retrospective sampling.

First, we assume that the data are obtained from prospective sampling, where the conditional density $f(\cdot;\boldsymbol{\theta})$ can be decomposed into

$$f(Y_i, \boldsymbol{M}_i|E_i;\boldsymbol{\theta}) = f_Y(Y_i|\boldsymbol{M}_i, E_i;\gamma_Y, \boldsymbol{\beta}_{FY}, \beta_{EY}, \psi^2_Y, \boldsymbol{\beta}_{EF}, \boldsymbol{\gamma}_M, \Lambda, \Psi^2) f_M(\boldsymbol{M}_i|E_i;\boldsymbol{\beta}_{EF}, \boldsymbol{\gamma}_M, \Lambda, \Psi^2),$$

where f_y belongs to

$$f_Y(Y_i;\zeta_i,\psi) = \exp\left[\{Y_i\zeta_i - b(\zeta_i)\}/a(\psi_Y) + c(Y_i,\psi_Y)\right] \tag{D.1}$$

with

$$\zeta_i = \gamma_Y + \beta_{EY}E_i + \boldsymbol{\beta}'_{FY}\boldsymbol{F}_i. \tag{D.2}$$

Next to show identifiability of Λ and Ψ^2, we note that for all values of Y, E and $\boldsymbol{M}$, if $f(\cdot;\boldsymbol{\theta}^1) = f(\cdot;\boldsymbol{\theta}^2)$ then $f(\boldsymbol{M}|E;\boldsymbol{\beta}^1_{EF}, \boldsymbol{\gamma}^1_M, \Lambda^1, \Psi^{2,1}) = f(\boldsymbol{M}|E;\boldsymbol{\beta}^2_{EF}, \boldsymbol{\gamma}^2_M, \Lambda^2, \Psi^{2,2})$. Based on the conditional distribution of $\boldsymbol{M}_i$ given E_i, which is multivariate normal with covariance matrix $\Lambda\Lambda' + \Psi$, and the sufficient condition 1 of Anderson and Rubin [2], we conclude that $\Psi^{2,1} = \Psi^{2,2}$ is identifiable and Λ is identifiable up to rotation by an orthogonal matrix (i.e. $\Lambda^1 Q = \Lambda^2$). To show identifiability of γ_M and $\boldsymbol{\beta}_{EF}$, we consider the conditional

mean of $\boldsymbol{M}_i$,

$$\mathbb{E}(\boldsymbol{M}|E) = \boldsymbol{\gamma}_M + \Lambda\boldsymbol{\beta}_{EF}E, \tag{D.3}$$

which implies $\Lambda^1\beta^1_{EF} = \Lambda^2\beta^2_{EF}$ and $\boldsymbol{\gamma}^1_M = \boldsymbol{\gamma}^2_M$. Because the matrices Λ^1 and Λ^2 have rank q, the solution to (D.3), $\boldsymbol{\beta}_{EF} = (\Lambda'\Lambda)^{-1}\Lambda'\,\mathbb{E}(\boldsymbol{M}_i|E_i)$ is unique up to orthogonal rotation.

Similarly, we show that for given sets of parameters Λ^1, $\boldsymbol{\gamma}^1_M$, $\boldsymbol{\beta}^1_{EF}$ and $\Lambda^2 = \Lambda^1 Q$, $\boldsymbol{\gamma}^2_M = \boldsymbol{\gamma}^1_M$, $\boldsymbol{\beta}^2_{EF} = Q'\boldsymbol{\beta}^2_{EF}$, equality of conditional densities $f_Y(Y|E,\boldsymbol{M};\gamma^1_Y,\boldsymbol{\beta}^1_{FY},\beta^1_{EY},\psi^1_Y) = f_Y(Y|E,\boldsymbol{M};\gamma^2_Y,\boldsymbol{\beta}^2_{FY},\beta^2_{EY},\psi^2_Y)$ implies equalities and $\boldsymbol{\beta}_{FY}$ up to orthogonal rotation (i.e. $Q'\boldsymbol{\beta}^1_{FY} = \boldsymbol{\beta}^2_{FY}$).

The density of Y belongs to the exponential family with $\zeta = \gamma_Y + \beta_{EY}E + \boldsymbol{\beta}'_{FY}\boldsymbol{F}$ and deviation parameter ψ_Y and has the form

$$f_Y(Y|E,\boldsymbol{M};\gamma_Y,\boldsymbol{\beta}_{FY},\beta_{EY},\psi_Y) = \int f_Y(Y|\zeta;\psi_Y)f_F(\boldsymbol{F}|E,\boldsymbol{M})d\boldsymbol{F}. \tag{D.4}$$

Conditional on E and $\boldsymbol{M}$, the factors $\boldsymbol{F}$ have a multivariate normal distribution with mean and covariance

$$\mathbb{E}(\boldsymbol{F}|E,\boldsymbol{M}) = \boldsymbol{\beta}_{FE}E + \Lambda'\left(\Lambda\Lambda' + \Psi^2\right)^{-1}(\boldsymbol{M} - \boldsymbol{\gamma}_M - \Lambda\boldsymbol{\beta}_{EF}E) = \boldsymbol{\gamma}^*_F + \boldsymbol{\beta}^*_{FE}E_i + \Lambda^{*'}\boldsymbol{M}_i, \tag{D.5}$$

$$Cov(\boldsymbol{F}|E,\boldsymbol{M}) = I - \Lambda'\left(\Lambda\Lambda' + \Psi^2\right)^{-1}\Lambda, \tag{D.6}$$

where $\boldsymbol{\gamma}^*_F = \Lambda'\left(\Lambda\Lambda' + \Psi^2\right)^{-1}\boldsymbol{\gamma}_M$, $\boldsymbol{\beta}^*_{FE} = (I - \Lambda'\left(\Lambda\Lambda' + \Psi^2\right)^{-1}\Lambda)\beta_{FE}$ and $\Lambda^* = \left(\Lambda\Lambda' + \Psi^2\right)^{-1}\Lambda$. For sets of parameters Λ^1, $\boldsymbol{\gamma}^1_M$, $\boldsymbol{\beta}^1_{EF}$ and $\Lambda^2 = \Lambda^1 Q$, $\boldsymbol{\gamma}^2_M = \boldsymbol{\gamma}^1_M$, $\boldsymbol{\beta}^2_{EF} = Q'\boldsymbol{\beta}^1_{EF}$, we observe $Q'\boldsymbol{\gamma}^{*1}_F = \boldsymbol{\gamma}^{*2}_F$, $Q'\boldsymbol{\beta}^{*1}_{FE} = \boldsymbol{\beta}^{*2}_{FE}$ and $\Lambda^{*1}Q = \Lambda^{*2}$.

Using (D.6), ζ can be written as $\zeta = \gamma_Y + \boldsymbol{\beta}'_{FY}\boldsymbol{\gamma}^*_F + \beta_{EY}E + \boldsymbol{\beta}'_{FY}\boldsymbol{\beta}^*_{FE}E + \boldsymbol{\beta}'_{FY}\Lambda^{*'}\boldsymbol{M} + \boldsymbol{\beta}'_{FY}\boldsymbol{Z}$ with random error coming from $\boldsymbol{Z} \sim N\left(0, I - \Lambda'\left(\Lambda\Lambda' + \Psi^2\right)^{-1}\Lambda\right)$. The integral in (D.4) is thus equivalent to

$$f_Y(Y|E,\boldsymbol{M};\gamma_Y,\boldsymbol{\beta}_{FY},\beta_{EY},\psi_y) = \int f_Y(Y|\zeta;\psi_Y)f_F(\boldsymbol{F}|E,\boldsymbol{M})d\boldsymbol{F} = \int f_Y(Y|\zeta;\psi_Y)f_z(\boldsymbol{Z})d\boldsymbol{Z}. \tag{D.7}$$

We now reformulated our set-up in a way that is equivalent to Berkson measurement error model [3]. In that model the parameters cannot be identified except when the variance of the term $\boldsymbol{Z}$ is known. Here, the variance of $\boldsymbol{Z}$ is , $I - \Lambda'\left(\Lambda\Lambda' + \Psi^2\right)^{-1}\Lambda$, is known. Therefore, identifiability of the fixed effects of E

and M, $\gamma_Y + \boldsymbol{\beta}'_{FY}\boldsymbol{\gamma}^*_F$, $\boldsymbol{\beta}'_{FY}\Lambda^{*'}$, $\beta_{EY} + \boldsymbol{\beta}'_{FY}\boldsymbol{\beta}^*_{EF}$ and the deviation parameter ψ_Y have been discussed in [4] and proven for linear and logistic regression models in [7]. As a result, we obtain

$$\gamma^1_Y + \boldsymbol{\beta}^{1'}_{FY}\boldsymbol{\gamma}^{*1}_F = \gamma^2_Y + \boldsymbol{\beta}^{2'}_{FY}\boldsymbol{\gamma}^{*2}_F,\ \beta^1_{EY} + \boldsymbol{\beta}^{1'}_{FY}\boldsymbol{\beta}^{*1}_{EF} = \beta^2_{EY} + \boldsymbol{\beta}^{2'}_{FY}\boldsymbol{\beta}^{*2}_{EF} \text{ and } \boldsymbol{\beta}^{1'}_{FY}\Lambda^{*1'} = \boldsymbol{\beta}^{2'}_{FY}\Lambda^{*2'} \tag{D.8}$$

Because $\Lambda^{*1}Q = \Lambda^{*2}$ and both matrices have rank q, the equality (D.8) implies $Q'\boldsymbol{\beta}^1_{FY} = \boldsymbol{\beta}^2_{FY}$ and remaining equalities $\gamma^1_Y = \gamma^2_Y$, $\beta^1_{EY} = \beta^2_{EY}$ and $\psi^1_y = \psi^2_y$ are straight forward to see. Hence, the proposition is proven for prospective sampling.

Under retrospective sampling, identifiability of the parameters $\boldsymbol{\theta}$ (except γ_Y) follows from the above proof and the form of the joint distribution of E and $\boldsymbol{M}$. Identifiability of γ_M, Λ, $\boldsymbol{\beta}_{EF}$ and Ψ^2 follows from the corresponding proofs for prospective sampling. Recall that the joint density of the observed data in controls, $f(E_i, M_i|Y_i = 0; \boldsymbol{\theta}, \gamma_E, \sigma^2_E)$ is a multivariate normal distribution with mean and covariance presented in section B.2. As result, identifiability of γ_E and σ^2_E from the marginal distribution of E in controls is trivial. Lastly, recall that in cases $f(E_i, M_i|Y_i = 0; \boldsymbol{\theta}, \gamma_E, \sigma^2_E)$ is multivariate normal distribution with the same covariance matrix but different means

$$\begin{bmatrix} \mu^1_E \\ \boldsymbol{\mu}^1_M \end{bmatrix} = \begin{bmatrix} \gamma_E \\ \boldsymbol{\gamma}_M + \Lambda\boldsymbol{\beta}_{EF}\gamma_E \end{bmatrix} + \begin{bmatrix} \sigma^2_E & \sigma^2_E\boldsymbol{\beta}'_{EF} \\ \sigma^2_E\Lambda\boldsymbol{\beta}_{EF} & \sigma^2_E\Lambda\boldsymbol{\beta}_{EF}\boldsymbol{\beta}'_{EF} + \Lambda \end{bmatrix} \begin{bmatrix} \beta_{EY} \\ \boldsymbol{\beta}_{FY} \end{bmatrix} \tag{D.9}$$

Identifiability of β_{EY} and $\boldsymbol{\beta}_{FY}$ up to orthogonal rotation follows from equality of mean vectors (D.9) between two joint distributions in cases. Since γ_E, $\gamma_M + \Lambda\boldsymbol{\beta}_{EF}\gamma_E$, σ^2_E and elements of the covariance matrix $\Sigma_{E,M}$ are unique, β_{EY} is unique and $\boldsymbol{\beta}_{FY}$ is unique up to orthogonal rotation.

*Proof of Proposition 2 (**Convergence of the EM algorithm**)*

Convergence of the EM algorithm to a single stationary point $\boldsymbol{\theta}^*_P$ follows from direct extension of Theorem 1 [12] to adaptive lasso penalties.

*Proof of Proposition 3 (**Consistency/Oracle-properties**)*

Let $A = \{j|\beta_{FY,j} \neq 0\}$ index the set of factors with effects on Y not equal to 0, $\hat{A} = \left\{j|\hat{\beta}_{FY,j} \neq 0\right\}$ index the set of factors with estimated effects on Y not equal to 0, $\hat{\boldsymbol{\beta}}^S_{FY} = \left\{\hat{\beta}_{FY,j} : j \in A\right\}$ be the vector of

estimates for the non-zero parameters and $\boldsymbol{\phi}^S = (\gamma_Y, \boldsymbol{\beta}_{FY}^S, \beta_{EY})$. Here, we show

$$\lim_{N\to\infty} P\left(\hat{A} = A\right) = 1 \tag{D.10}$$

and as $N \to \infty$

$$\sqrt{N}(\hat{\boldsymbol{\phi}}^S - \boldsymbol{\phi}^S) \to_d N\left(0, \mathbb{I}^{-1}_{\boldsymbol{\phi}^S}\right), \tag{D.11}$$

for both prospectively and retrospectively sampled data.

Consistency and asymptotic normality for other parameters, $\hat{\gamma}_E$, $\hat{\boldsymbol{\gamma}}_M$, the factor loadings $\hat{\Lambda}^S$ and effects $\hat{\boldsymbol{\beta}}_{FE}^S$ follow from the decomposition of the distribution f into four parts as in (10), Theorem 4.2 of Shrivastava et al. (2014) and the proofs for (D.10) and (D.11).

We prove statements (D.10) and (D.11) by considering one iteration of the EM algorithm, which is a variation of M-estimation (Van der Vaart, 2000). Let $\hat{\boldsymbol{\theta}}_0$ be the initial $\sqrt{N}$-consistent estimate of $\boldsymbol{\theta} = (\gamma_Y, \boldsymbol{\beta}_{FY}, \beta_{EY}, \sigma_Y^2, \gamma_E, \boldsymbol{\beta}_{EF}, \sigma_E^2, \boldsymbol{\gamma}_M, \Lambda, \Psi^2)$. Let $\gamma_Y^*, \boldsymbol{\beta}_{FY}^*, \beta_{EY}^*$ and A^* be the true values, $\gamma_Y = \gamma_Y^* + \frac{u_{\gamma_Y}}{\sqrt{N}}$, $\boldsymbol{\beta}_{FY} = \boldsymbol{\beta}_{FY}^* + \frac{\boldsymbol{u}_{\beta_{FY}}}{\sqrt{N}}$, $\beta_{EY} = \beta_{EY}^* + \frac{u_{\beta_{EY}}}{\sqrt{N}}$ and for convenience we set $\boldsymbol{\beta}^* = (\gamma_Y^*, \beta_{EY}^*, \boldsymbol{\beta}_{FY}^*)$, $\boldsymbol{u} = (u_{\gamma_Y}, u_{\beta_{EY}}, \boldsymbol{u}_{\beta_{FY}})$ and $\boldsymbol{X}_i = (1, E_i, \boldsymbol{F}_i)$. Then, the expected penalized full log-likelihood with a fixed value of ψ_Y is

$$EPFLL_n(u_{\gamma_Y}, \boldsymbol{u}_{\beta_{FY}}, u_{\beta_{EY}}) = \sum_{i=1}^{N} \mathbb{E}\left[Y_i(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i + \frac{\boldsymbol{u}'\boldsymbol{X}_i}{\sqrt{N}}) - b(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i + \frac{\boldsymbol{u}'\boldsymbol{X}_i}{\sqrt{N}})|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right] + \rho_n \sum_{j=1}^{q} \frac{|\beta_{FY,j}^* + \frac{\boldsymbol{u}_{\beta_{FY},j}}{\sqrt{N}}|}{|\hat{\beta}_{FY,j}^0|}. \tag{D.12}$$

Let $(\hat{u}_{\gamma_Y}, \hat{\boldsymbol{u}}_{\beta_{FY}}, \hat{u}_{\beta_{EY}}) = \arg\min_{u_{\gamma_Y}, \boldsymbol{u}_{\beta_{FY}}, u_{\beta_{EY}}} EPFLL_n(u_{\beta_0}, \boldsymbol{u}_{\beta_{FY}}, u_{\beta_{EY}}) - EPFLL_n(0, \mathbf{0}, 0)$; then one step-estimates for γ_Y, $\boldsymbol{\beta}_{FY}$ and β_{EY} are $\hat{\gamma}_Y = \gamma_Y^* + \frac{\hat{u}_{\gamma_Y}}{\sqrt{N}}$, $\hat{\boldsymbol{\beta}}_{FY} = \boldsymbol{\beta}_{FY}^* + \frac{\hat{\boldsymbol{u}}_{\beta_{FY}}}{\sqrt{N}}$ and $\hat{\beta}_{EY} = \beta_{EY}^* + \frac{\hat{u}_{\beta_{EY}}}{\sqrt{N}}$. Using a

second order Taylor expansion,

$$EPFLL_n(u_{\gamma_Y}, \boldsymbol{u}_{\beta_{FY}}, u_{\beta_{EY}}) - EPFLL_n(0, \mathbf{0}, 0) \sim \underbrace{\sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - b'(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i)\right)\frac{\boldsymbol{X}_i'\boldsymbol{u}}{\sqrt{N}}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right]}_{\text{Part I}}$$
$$\underbrace{-\sum_{i=1}^{N} \mathbb{E}\left[\frac{b''(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i)\boldsymbol{u}'\boldsymbol{X}_i\boldsymbol{X}_i'\boldsymbol{u}}{2N}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right]}_{\text{Part II}} \underbrace{-\sum_{i=1}^{N} \mathbb{E}\left[\frac{b'''(\boldsymbol{\beta}^{+'}\boldsymbol{X}_i)(\boldsymbol{u}'\boldsymbol{X}_i)^3}{6N^{3/2}}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right]}_{\text{Part III}}$$
$$\underbrace{+\frac{\rho_n}{\sqrt{N}}\sum_{j=1}^{q}\sqrt{N}\frac{|\beta^*_{FY,i} + \frac{u_{\beta_{FY,j}}}{\sqrt{N}}| - |\beta^*_{FY,j}|}{|\hat{\beta^0_{FY,j}}|}}_{\text{Part IV}} \quad \text{(D.13)}$$

where $\boldsymbol{\beta}^+$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + \frac{\boldsymbol{u}}{\sqrt{N}}$.

The sum in Part I converges to $\boldsymbol{u}'\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}$ following a multivariate normal distribution with mean $\mathbf{0}$ and covariance

$$cov(\boldsymbol{\epsilon}) = \mathbb{E}\left[\left(Y_i - b'(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i)\right)^2 \boldsymbol{X}_i'\boldsymbol{X}; \boldsymbol{\theta}^*\right] = \mathbb{I}_{\boldsymbol{\phi}^S},$$

where $\mathbb{I}_{\boldsymbol{\phi}^S}$ is the Fisher information matrix. By the properties of the second derivative of a density belonging to the GLM class, the Part II of the sum converges to

$$\lim_{N\to\infty}\sum_{i=1}^{N}\mathbb{E}\left[\frac{b''(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i)\boldsymbol{u}'\boldsymbol{X}_i\boldsymbol{X}_i'\boldsymbol{u}}{N}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right] = \boldsymbol{u}'\,\mathbb{E}\left[\left(Y_i - b'(\boldsymbol{\beta}^{*'}\boldsymbol{X}_i)\right)^2 \boldsymbol{X}_i\boldsymbol{X}_i'; \boldsymbol{\theta}^*\right]\boldsymbol{u} = \boldsymbol{u}'\,\mathbb{I}_{\boldsymbol{\phi}^S}\,\boldsymbol{u}$$

Because $b'''(x) < 5$ for any $\boldsymbol{X}$ and $\boldsymbol{X}$ have finite moments, the Part III the sum in D.13 converges in probability to

$$\lim_{N\to\infty}\sum_{i=1}^{N}\mathbb{E}\left[\frac{b'''(\boldsymbol{\beta}^{+'}\boldsymbol{X}_i)(\boldsymbol{u}'\boldsymbol{X}_i)^3}{N^{3/2}}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right] = 0.$$

Lastly, we derive the convergence of Part IV in (D.13) using assumption that $\rho_N/\sqrt{N} \to 0$ and $\rho_N \to \infty$. Let $\beta^*_{FY,j} \neq 0$, then

$$\frac{\rho_n}{\sqrt{N}}\sqrt{N}\frac{|\beta^*_{FY,j} + \frac{u_{\beta_{FY,j}}}{\sqrt{N}}| - |\beta^*_{FY,j}|}{|\hat{\beta^0_{FY,j}}|} = \frac{\rho_n}{\sqrt{N}}\frac{1}{|\hat{\beta}^0_{FY,j}|}u_{\alpha,i}sign(\beta^*_{FY,j}) \to_p 0 \text{ by Slutsky's theorem.} \quad \text{(D.14)}$$

Let $\beta^*_{FY,j} = 0$, then

$$\frac{\rho_n}{\sqrt{N}}\sqrt{N}\frac{|\beta^*_{FY,j} + \frac{u_{\beta_{FY,j}}}{\sqrt{N}}| - |\beta^*_{FY,j}|}{|\hat{\beta^0_{FY,j}}|} = \rho_n \frac{1}{|\sqrt{N}\hat{\beta}^0_{FY,j}|} u_{\alpha,j} \to_d \begin{cases} 0 & \text{if } u_{\beta_{FY,j}} = 0 \\ \infty & \text{if } u_{\beta_{FY,j}} \neq 0 \end{cases} \text{ by Slutsky's theorem.} \tag{D.15}$$

By combining (D.14) and (D.15), we obtain the following limiting convergence of *Part IV*

$$\frac{\rho_n}{\sqrt{N}}\sqrt{N}\frac{|\beta^*_{FY,j} + \frac{u_{\beta_{FY,j}}}{\sqrt{N}}| - |\beta^*_{FY,j}|}{|\hat{\beta}^0_{FY,j}|} \to_d \begin{cases} 0 & \text{if } u_{\beta_{FY,j}} = 0 \text{ for all } j \notin A^* \\ \infty & \text{otherwise.} \end{cases} \tag{D.16}$$

In conclusion, we observed that

$$EPFLL_n(u_{\gamma_Y}, \boldsymbol{u}_{\beta_{FY}}, u_{\beta_{EY}}) - EPFLL_n(0, \mathbf{0}, 0) \to_d \begin{cases} \boldsymbol{u'}_{\boldsymbol{\phi}^S} \boldsymbol{\epsilon}_{\boldsymbol{\phi}^S} - 1/2 \boldsymbol{u'}_{\boldsymbol{\phi}^S} \mathbb{I}_{\boldsymbol{\phi}^S} \boldsymbol{u}_{\boldsymbol{\phi}^S} & \text{if } u_{\beta_{FY,j}} = 0 \text{ for all } j \notin A^* \\ \infty & \text{otherwise,} \end{cases} \tag{D.17}$$

where $\epsilon_{\boldsymbol{\phi}^S} = (\epsilon_{\gamma_Y}, \epsilon_{\beta_{FY},A^*}, \epsilon_{\beta_{EY}})$, $u_{\boldsymbol{\phi}^S} = (u_{\gamma_Y}, u_{\beta_{FY},A^*}, u_{\beta_{EY}})$ and $I_{\boldsymbol{\phi}^S}$ is Fisher' information for γ_Y, $\boldsymbol{\beta}^{S*}_{FY}$ and β_{EY}. The difference, $EPFLL(u_{\gamma_Y}, \boldsymbol{u}_{\beta_{FY}}, u_{\beta_{EY}}) - EPFLL(0, \mathbf{0}, 0)$, is convex and the unique minimum of (D.17) is

$$(\hat{u}_{\gamma_Y}, \hat{u}_{\beta_{EY}}, \hat{\boldsymbol{u}}_{\beta_{FY},A^*}, \hat{\boldsymbol{u}}_{\beta_{FY},A^{*c}})^\infty = \left(\mathbb{I}^{-1}_{\boldsymbol{\phi}^S} \boldsymbol{\epsilon}_{\boldsymbol{\phi}^S}, \mathbf{0} \right). \tag{D.18}$$

Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000); we have

$$(\hat{u}_{\beta_0}, \hat{u}_{\beta_{EY}} \hat{\boldsymbol{u}}_{\beta_{FY},A^*}, \hat{\boldsymbol{u}}_{\beta_{FY},A^{*c}}) \to_d \left(\mathbb{I}^{-1}_{\boldsymbol{\phi}^S} \boldsymbol{\epsilon}_{\boldsymbol{\phi}^S}, \mathbf{0} \right) \tag{D.19}$$

Therefore, we prove asymptotically normality.

Next we show the consistency part. For any $j \in A^*$, the asymptotic normality results indicate that $\hat{\beta}_{FY,j} \to_p \beta^*_{FY,j}$; thus $P(j \in \hat{A}) \to_p 1$. It is left to show that for any $j' \notin A^*$, $P(j' \in \hat{A}) \to_p 0$. By the

Karush-Kuhn-Tucker optimality conditions, we have

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - b'(\hat{\boldsymbol{\beta}}\boldsymbol{X}_i)\right) F_{j'}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right] = \rho_n \frac{1}{|\hat{\beta}^0_{FY,j'}|} \tag{D.20}$$

By dividing both parts by $\sqrt{N}$ and using normal convergence of the estimates ($\hat{\boldsymbol{\beta}}$), square-root consistent initial estimates ($\hat{\boldsymbol{\theta}}_0$) and Slutsky's theorem (Delta method), we observe

$$\frac{\sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - b'(\hat{\boldsymbol{\beta}}\boldsymbol{X}_i)\right) F_{j'}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right]}{\sqrt{N}} \to_d N(0, \sigma_s^2), \tag{D.21}$$

and

$$\frac{\rho_n}{\sqrt{N}|\hat{\beta}^0_{FY,j'}|} \to_p \infty. \tag{D.22}$$

Thus,

$$P\left(j' \in \hat{A}_{\boldsymbol{\beta}_{FY}}\right) \leq P\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - b'(\hat{\boldsymbol{\beta}}\boldsymbol{X}_i)\right) F_{j'}|Y_i, E_i, \boldsymbol{M}_i; \hat{\boldsymbol{\theta}}_0\right] = \rho_n \frac{1}{\sqrt{N}|\hat{\beta}^0_{FY,j'}|}\right) \to 0 \tag{D.23}$$

Consistency and asymptotic normality of $\hat{\psi}_Y$ follow from the estimation equations, asymptotic normality of the regression coefficients and Slutsky's theorem.

The proof for the retrospective likelihood follows from similar arguments. Here we again do not show consistency and asymptotic normality for $\hat{\boldsymbol{\gamma}}_M$, factor loadings $\hat{\Lambda}$ and $\hat{\Psi}^2$ because of the decomposition of the likelihood (16). Here we demonstrate consistency and asymptotic normality for other parameters from the observation that $\gamma_E, \boldsymbol{\beta}_{FY}, \beta_{EY}, \boldsymbol{\beta}_{EF}$ are estimated from (C.5) with σ_E^2 fixed at $\hat{\sigma}_E^{20}$ initial square root N estimate, which can be expanded to

$$\underset{\gamma_E,\boldsymbol{\beta}_{FY},\beta_{EY},\boldsymbol{\beta}_{EF}}{\arg\max}\left[\sum_{i=1}^{N_1}-\mathbb{E}\left(\sum_{j=1}^{q}\frac{(F_i[j]-\boldsymbol{\beta}_{EF}[j]E_i-\boldsymbol{\beta}_{FY}[j])^2}{2}|Y_i=1,E_i,\boldsymbol{M}_i;\hat{\boldsymbol{\theta}}_0\right)\right.$$
$$-\sum_{i=1}^{N_0}\mathbb{E}\left(\sum_{j=1}^{q}\frac{(F_i[j]-\boldsymbol{\beta}_{EF}[j]E_i)^2}{2}|Y_i=0,E_i,\boldsymbol{M}_i;\hat{\boldsymbol{\theta}}_0\right)-\sum_{i=1}^{N_1}\frac{(E_i-\gamma_E-\hat{\sigma}_E^{20}\boldsymbol{\beta}'_{EF}\boldsymbol{\beta}_{FY}-\hat{\sigma}_E^{20}\beta_{EY})^2}{2\hat{\sigma}_E^{20}}$$
$$\left.-\sum_{i=1}^{N_0}\frac{(E_i-\gamma_E)^2}{2\hat{\sigma}_E^{20}}-\rho_1\sum_{j=1}^{q}\frac{|\beta_{FY,j}|}{|\hat{\beta}^0_{FY,j}|}-\rho_2\sum_{j=1}^{q}\frac{|\beta_{EF,j}|}{|\hat{\beta}^0_{EF,j}|}\right]. \quad \text{(D.24)}$$

Consistency and asymptotic normality for these parameters follows directly form of (D.24) and the proof for prospective sampling. These are two penalized linear regressions with one of two parameters in $\hat{\sigma}_E^{20}\boldsymbol{\beta}'_{EF}\boldsymbol{\beta}_{FY}$ is treated as covariate. In this case, $\mathbb{I}_{\phi^S}$ is information matrix for $\boldsymbol{\phi}^S=(\gamma_Y,\boldsymbol{\beta}^S_{FY},\beta_{EY},\boldsymbol{\beta}^S_{EF})$ This completes the proof of proposition 3.

Web Appendix E Covariate adjustment

Here, we rewrite our model, originally described by equations (4-7), to include a covariate vector $\boldsymbol{X}_i=(X_{iq},...,X_{il})'$:

$$f(Y_i;\zeta_i,\psi)=\exp\left[\{Y_i\zeta_i-b(\zeta_i)\}/a(\psi_Y)+c(Y_i,\psi_Y)\right] \quad \text{(E.1)}$$

with

$$\zeta_i=\gamma_Y+\beta_{EY}E_i+\boldsymbol{\beta}'_{FY}\boldsymbol{F}_i+\boldsymbol{\beta}'_{XY}\boldsymbol{X}_i, \quad \text{(E.2)}$$

$$\boldsymbol{F}_i=\boldsymbol{\beta}_{EF}E_i+\Omega_{XF}\boldsymbol{X}_i+\boldsymbol{e}_f \text{ with } \boldsymbol{e}_f\sim N(0,I_q), \quad \text{(E.3)}$$

and

$$\boldsymbol{M}_i=\boldsymbol{\gamma}_M+\Lambda\boldsymbol{F}_i+\boldsymbol{e}_m \text{ with } e_m\sim N(0,\Psi^2),\Psi^2=diag\{\psi_1^2,...,\psi_p^2\}, \quad \text{(E.4)}$$

where $\boldsymbol{\beta}'_{XY}=(\beta_{XY,1},...,\beta'_{XY,l})^T$ is a vector of length l and Ω_{XF} is a q by l matrix. To estimate the vector of parameters $\boldsymbol{\theta}=(\gamma_Y,\boldsymbol{\beta}_{FY},\beta_{EY},\boldsymbol{\beta}_{XY},\psi_Y,\boldsymbol{\beta}_{EF},\Omega_{XF},\boldsymbol{\gamma}_M,\Lambda,\Psi^2)$ using the observed data, $(Y_i,\boldsymbol{M}_i,E_i,\boldsymbol{X}_i)$ for $i=1,...,N$, we derive prospective and retrospective likelihoods.

E.1 Prospective likelihood

Under prospective sampling, the full data likelihood for $(Y, \boldsymbol{M}, \boldsymbol{F})$ is

$$L_P^F(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i | E_i, \boldsymbol{X}_i; \boldsymbol{\theta}), \tag{E.5}$$

where f is the product of the densities defined by equations (E.1-E.4),

$$f(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i | E_i, \boldsymbol{X}_i; \boldsymbol{\theta}) = f_Y(Y_i | \boldsymbol{F}_i, E_i, \boldsymbol{X}_i; \gamma_Y, \boldsymbol{\beta}_{FY}, \boldsymbol{\beta}_{XY}, \beta_{EY}, \psi_Y) f_M(\boldsymbol{M}_i | \boldsymbol{F}_i; \boldsymbol{\gamma}_M, \Lambda, \Psi^2) f_F(\boldsymbol{F}_i | E_i, \boldsymbol{X}_i; \boldsymbol{\beta}_{EF}, \Omega_{XF}). \tag{E.6}$$

Here, we use f_M, f_Y, and f_F to denote the implied distribution of $\boldsymbol{M}$, Y, and $\boldsymbol{F}$, respectively. However, the factors $\boldsymbol{F}_i$ are not observed. The likelihood for the observed data, $(Y_i, \boldsymbol{M}_i)$, is therefore

$$L_P^O(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int_{\boldsymbol{F}} f(Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i | E_i, \boldsymbol{X}_i; \boldsymbol{\theta}) d\boldsymbol{F}. \tag{E.7}$$

Although $L_P^O(\boldsymbol{\theta})$ does not have a closed form in general, $L_P^O(\boldsymbol{\theta})$ can be written as the product of normal distributions when Y_i is normally distributed. Based on the model in (E.1-E.4), the conditional distribution of $Y_i, \boldsymbol{M}_i, \boldsymbol{F}_i$ in (E.5) is multivariate normal with mean vector $(\mu_Y, \boldsymbol{\mu}_M, \boldsymbol{\mu}_F)$,

$$\mu_Y = \gamma_Y + (\boldsymbol{\beta}'_{FY} \boldsymbol{\beta}_{EF} + \beta_{EY}) E_i + (\boldsymbol{\beta}'_{FY} \Omega_{XF} + \boldsymbol{\beta}'_{XY}) \boldsymbol{X}_i,$$

$$\boldsymbol{\mu}_M = \boldsymbol{\gamma}_M + \Lambda \boldsymbol{\beta}_{EF} E_i + \Lambda \Omega_{XF} \boldsymbol{X}_i,$$

$$\boldsymbol{\mu}_F = \boldsymbol{\beta}_{EF} E_i + \Omega_{XF} \boldsymbol{X}_i,$$

and the same covariance matrices (B.2-B.4). Identifiability of all parameters follows from *Proposition 1* while treating exposure E as a vector (e.g $(E_i, \boldsymbol{X}_i)$). We can estimate the parameters using a similar EM procedure to that initially described for the scenario without covariates (see Web Appendix Web Appendix C).

E.2 Retrospective likelihood

Under retrospective sampling, N_1 cases and N_0 controls are drawn from the population of cases and controls, respectively, and biomarkers, exposures and covariates are observed ($N_1 + N_0 = N$). Here we assume E_i conditional on $\boldsymbol{X}_i$ to be normally distributed with mean $\gamma_E + \boldsymbol{\beta}'_{XE}\boldsymbol{X}_i$ and variance σ_E^2 in the overall population. The corresponding likelihood has the form

$$L_R^F(\boldsymbol{\theta}) = \prod_{i \in case} f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | Y_i = 1, \boldsymbol{X}_i; \boldsymbol{\theta}) \prod_{i \in control} f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | Y_i = 0, \boldsymbol{X}_i; \boldsymbol{\theta}). \tag{E.8}$$

To approximate the likelihood, we assume the outcome is rare in the general population, i.e.,

$$P(Y_i = 1 | E_i, \boldsymbol{F}_i, \boldsymbol{X}_i; \gamma_Y, \beta_{EY}, \boldsymbol{\beta}_{FY}, \boldsymbol{\beta}_{XF}) = \frac{\exp(\gamma_Y + \beta_{EY}E_i + \boldsymbol{\beta}'_{FY}\boldsymbol{F}_i + \boldsymbol{\beta}'_{XY}\boldsymbol{X}_i)}{1 + \exp(\gamma_Y + \beta_{EY}E_i + \boldsymbol{\beta}'_{FY}\boldsymbol{F}_i + \boldsymbol{\beta}'_{XY}\boldsymbol{X}_i)} \approx \exp(\gamma_Y + \beta_{EY}E_i + \boldsymbol{\beta}'_{FY}\boldsymbol{F}_i + \boldsymbol{\beta}'_{XY}\boldsymbol{X}_i). \tag{E.9}$$

Under the rare disease assumption, the distribution of $(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i)$ in controls is approximately equal to the distribution in the general population. Thus, under models (E.1-E.4)

$$f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | Y_i = 0, \boldsymbol{X}_i; \boldsymbol{\theta}) \approx f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | \boldsymbol{X}_i; \boldsymbol{\theta}) = \phi\left(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | \boldsymbol{X}_i; \boldsymbol{\mu}_{0|X}, \Sigma_{E,M,F}\right), \tag{E.10}$$

where $\phi(\cdot | \boldsymbol{X}_i; \boldsymbol{\mu}_{0|X}, \Sigma_{E,M,F})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_{0|X} = (\mu_{E|X}, \boldsymbol{\mu}_{M|X}, \boldsymbol{\mu}_{F|X})$ and the covariance matrix $\Sigma_{E,M,F}$. Under models (E.1-E.4) and the distributional assumption about E_i, the mean vectors are

$$\boldsymbol{\mu}_{M|X} = \boldsymbol{\gamma}_M + \Lambda\left(\boldsymbol{\beta}_{EF}\gamma_E + (\Omega_{XF} + \boldsymbol{\beta}_{EF}\boldsymbol{\beta}'_{XE})\boldsymbol{X}_i\right),$$

$$\boldsymbol{\mu}_{F|X} = \boldsymbol{\beta}_{EF}\gamma_E + (\Omega_{XF} + \boldsymbol{\beta}_{EF}\boldsymbol{\beta}'_{XE})\boldsymbol{X}_i,$$

$$\mu_{E|X} = \gamma_E + \boldsymbol{\beta}'_{XE}\boldsymbol{X}_i,$$

and the covariance matrices in $\Sigma_{E,M,F}$ are equal to (B.6-B.8).

The distribution of $(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i)$ in cases under model (E.9) is

$$f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | Y_i = 1, \boldsymbol{X}_i; \boldsymbol{\theta}) = \frac{f(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i, Y_i = 1 | \boldsymbol{X}_i; \boldsymbol{\theta})}{P(Y_i = 1 | \boldsymbol{X}_i; \boldsymbol{\theta})} \approx \phi\left(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i | \boldsymbol{X}_i; \boldsymbol{\mu}_{1|X}, \Sigma_{E,M,F}\right) \tag{E.11}$$

where $\boldsymbol{\mu}_{1|X} = (\mu^1_{E|X}, \boldsymbol{\mu}^1_{M|X}, \boldsymbol{\mu}^1_{F|X})' = (\mu_{E|X}, \boldsymbol{\mu}_{M|X}, \boldsymbol{\mu}_{F|X})' + \Sigma_{E,M,F}(\beta_{EY}, \mathbf{0}', \boldsymbol{\beta}'_{FY})'$. We note that the effects of $\boldsymbol{X}_i$ on Y_i (i.e. $\boldsymbol{\beta}_{XY}$) are not included in $\boldsymbol{\mu}_{1|X}$ because those parameters appear in both the numerator and denominator of (E.11) and cancel out. Based on the above approximations, the likelihood for the full data $(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i)$ is

$$L^F_R(\boldsymbol{\theta}) = \prod_{i \in case} \phi(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\mu}_{1|X}, \Sigma_{E,M,F}) \prod_{i \in controls} \phi(E_i, \boldsymbol{M}_i, \boldsymbol{F}_i; \boldsymbol{\mu}_{0|X}, \Sigma_{E,M,F}) \tag{E.12}$$

and the likelihood for the observed data is therefore easily shown to be

$$L^O_R(\boldsymbol{\theta}) = \prod_{i \in case} \phi(E_i, \boldsymbol{M}_i; \boldsymbol{\mu}_{1;M,E|X}, \Sigma_{M,E}) \prod_{i \in controls} \phi(E_i, \boldsymbol{M}_i; \boldsymbol{\mu}_{0;M,E|X}, \Sigma_{M,E}),, \tag{E.13}$$

where $\mu_{1;E,M|X} = (\mu^1_{E|X}, \boldsymbol{\mu}^1_{M|X})$, $\mu_{0;E,M|X} = (\mu_{E|X}, \boldsymbol{\mu}_{M|X})$, and $\Sigma_{E,M}$ is the appropriate sub-matrix of $\Sigma_{E,M,F}$. Under the retrospective likelihoods (E.13 or E.12), both the intercept γ_Y and vector of effects $\boldsymbol{\beta}_{XY}$ are not identifiable. Identifiability of the other parameters follows from Proposition 1 and the form of the mean vectors. We can estimate the parameters using a similar EM procedure to that initially described for the scenario without covariates (Web Appendix C).

E.2.1 Matched case-control design

Here, we derive two methods for analyzing case-control data. For simplicity of exposition, we shall assume "one-to-one" matching for each case/control pair and we will let $E^1_i, \boldsymbol{M}^1_i, \boldsymbol{F}^1_i, Y^1_i, \boldsymbol{X}^1_i$ be a vector random variables for the case and $E^2_i, \boldsymbol{M}^2_i, \boldsymbol{F}^2_i, Y^2_i, \boldsymbol{X}^2_i$ be a vector random variables for the matched control. We shall assume that we first identify a case and then, based on the matching variables $\boldsymbol{X}_i$, select the control. The first method conditions on there being one case per pair [9] and describes the full likelihood by

$$L^F_{MCC}(\boldsymbol{\theta}) = \prod_{i=1}^{N_1} f(Y^1_i, \boldsymbol{M}^1_i, \boldsymbol{F}^1_i, E^1_i, Y^2_i, \boldsymbol{M}^2_i, \boldsymbol{F}^2_i, E^2_i | \boldsymbol{X}^1_i, \boldsymbol{X}^2_i, Y^1_i + Y^2_i = 1, \boldsymbol{X}^1_i = \boldsymbol{X}^2_i; \boldsymbol{\theta}), \tag{E.14}$$

which can be rewritten as

$$L^F_{MCC}(\boldsymbol{\theta}) = \prod_{i=1}^{N_1} \frac{f(Y_i^1 = 1, \boldsymbol{M}_i^1, \boldsymbol{F}_i^1, E_i^1, Y_i^2 = 0, \boldsymbol{M}_i^2, \boldsymbol{F}_i^2, E_i^2 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, Y_i^1 + Y_i^2 = 1, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta})}{P(Y_i^1 + Y_i^2 = 1 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta})} \quad \text{(E.15)}$$

Under the assumption of a rare outcome

$$P(Y_i^1 + Y_i^2 = 1 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta}) \approx P(Y_i^1 = 1 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta})$$
$$+ P(Y_i^2 = 1 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta}) = 2P(Y_i^1 = 1 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta}),$$

which leads to a likelihood equivalent to that of a retrospective study without matching

$$L^F_{MCC}(\boldsymbol{\theta}) = \prod_{i=1}^{N_1} \frac{f(Y_i^1 = 1, \boldsymbol{M}_i^1, \boldsymbol{F}_i^1, E_i^1 | \boldsymbol{X}_i^1; \boldsymbol{\theta}) f(Y_i^2 = 0, \boldsymbol{M}_i^2, \boldsymbol{F}_i^2, E_i^2 | \boldsymbol{X}_i^2; \boldsymbol{\theta}) I(\boldsymbol{X}_i^1 = \boldsymbol{X}_i^2)}{2P(Y_i^1 = 1 | \boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \boldsymbol{X}_i^1 = \boldsymbol{X}_i^2; \boldsymbol{\theta})}. \quad \text{(E.16)}$$

$$\approx \prod_{i=1}^{N_1} \phi(E_i^1, \boldsymbol{M}_i^1, \boldsymbol{F}_i^1; \boldsymbol{\mu}_{1|X}, \Sigma_{E,M,F}) \phi(E_i^2, \boldsymbol{M}_i^2, \boldsymbol{F}_i^2; \boldsymbol{\mu}_{0|X}, \Sigma_{E,M,F}). \quad \text{(E.17)}$$

Hence, inference on (E.14) is the same as for the retrospective likelihood presented in the previous section. We note, as discussed by VanderWeele and Tchetgen Tchetgen [13], our likelihood still includes covariates, $\boldsymbol{X}_i$, because we need to estimate the effect of the exposure on the mediating factors in the presence of these covariates. The second method for matched case-control study designs takes the difference between the vector of measurements from the first and second observation within a pair. This procedure removes the nuisance effects of $\boldsymbol{X}$,$\boldsymbol{\beta}_{XF}$ and $\boldsymbol{\beta}_{XE}$ from the estimation. The vector of differences $(\delta E_i, \delta \boldsymbol{M}_i, \delta \boldsymbol{F}_i)' = (E_i^1, \boldsymbol{M}_i^1, \boldsymbol{F}_i^1)' - (E_i^2, \boldsymbol{M}_i^2, \boldsymbol{F}_i^2)'$ is distributed as

$$(\delta \boldsymbol{M}_i, \delta \boldsymbol{F}_i, \delta E_i)' \sim N\left(\Sigma_{E,M,F}(\beta_{EY}, \boldsymbol{0}', \boldsymbol{\beta}'_{FY})', 2\Sigma_{E,M,F}\right), \quad \text{(E.18)}$$

and does not contain $\boldsymbol{\beta}_{XF}$ and $\boldsymbol{\beta}_{XE}$. This approach eliminates the effects of covariates $\boldsymbol{X}_i$; however, we prefer the former method for simplicity of description and the potential loss of precision from the reduction of degrees of freedom.

Web Appendix F Performance of Procedure: Violations in Model Assumptions and Alternative Criteria for Model Selection

We evaluated how our procedure perform when some of the underlying assumptions are violated. Specifically, we considered scenarios where (i) the exposure directly affected the biomarkers (ii) the biomarkers were not normally distributed (iii) the number of factors was misspecified. We also evaluated how our procedure would perform if we used alternative criteria for model selection (e.g. BIC, AIC instead of EBIC). We evaluated performance by simulations.

F.1 Simulations: Set-up

We first evaluated the properties of the procedures when the exposure directly affects individual biomarkers and those biomarkers, in turn, directly affect the outcome (see Web Figure 1). Our underlying model could then be described by:

$$\zeta_i = \gamma_Y + \beta_{EY} E_i + \sum_{j=1}^{20} \beta^D_{MY,j} M_{ij}, \tag{F.1}$$

and for these first 20 biomarkers

$$M_{ij} = \beta^D_{EM,j} E_i + \boldsymbol{e}_{m,i} \text{ with } e_{m,i,j} \sim N(0,1). \tag{F.2}$$

We also assumed that $\boldsymbol{F}_i$ and $\boldsymbol{M}_i$ are normally distributed:

$$\boldsymbol{F}_i = \boldsymbol{\beta}_{EF} E_i + \boldsymbol{e}_{f,i} \text{ with } \boldsymbol{e}_{f,i} \sim N(0, I_q), \tag{F.3}$$

where I_q is the q by q identity matrix and for other biomarkers

$$\boldsymbol{M}_i = \boldsymbol{\gamma}_M + \Lambda \boldsymbol{F}_i + \boldsymbol{e}_{m,i} \text{ with } \boldsymbol{e}_{m,i} \sim N(0, \Psi^2), \Psi^2 = diag\{\psi_1^2, ..., \psi_p^2\}. \tag{F.4}$$

For these simulations, the first twenty biomarkers were mediators (see Web Figure 1; $\beta^D_{EM,j} \in \{0.4 \cdot 0.3 = 0.12, 0.5 \cdot 0.3 = 0.15\}$, $\beta^D_{MY,j} \in \{0.4 \cdot 0.3 = 0.12, 0.5 \cdot 0.3 = 0.15\}$ for $j = 1, ..., 20$). Moreover, we assumed that there were fourteen factors ($q = 14$), each affecting twenty unique biomarkers; the first four factors were associated only with E and the last ten factors were associated with neither E nor Y. Otherwise parameters were similar to those described in the main text: $\lambda_1 \in \{0.25, 0.3\}$, $\lambda_0 \in \{0.4, 0.5\}$, $\beta_{EY} = 0.3$, $\beta_{EF,1} = ... = \beta_{EF,4} = 0.7$, $\gamma_{M,1} = ... = \gamma_{M,p} = 0.5$, $\psi^2_1 = ... = \psi^2_p = 1$, $\gamma_Y = -4.6$ (i.e. $P(Y = 1|E = 0, F_1 = 0) = 0.01$ for binary outcomes), and $E \sim N(0.5, 1)$.

We next evaluated the properties of the procedures when the biomarkers were not normally distributed. Specifically, we considered two scenarios. For scenario 1, we assumed that the random error (i.e. $e_{m,i}$ in equation 4 of the main text) followed a normalized chi-squared distribution: $e_{m,i} \sim (1/\sqrt{100})\chi^2(df = 50) - 50/\sqrt{100}$. For scenario 2, we assumed that the random error followed a standard normal distribution, but the values were truncated at -1.7 (The $\approx 5\%$ of values below -1.7 were set to -1.7.). All other parameters were identical those described in the main text.

We next evaluated the properties of the procedures when we assumed that number of factors was 20 while, in truth, $q = 15$. Finally, we compared the properties of the procedure when using AIC or BIC instead of EBIC.

F.2 Simulations: Results

We describe the properties of the procedures when the biomarkers are directly related to the exposure and the outcome biomarkers in Web Figures 2 - 4. The key result is that, for all scenarios, IMA outperforms LMVA, TMAO and TMAR. In most scenarios, LVMA, TMAO and TMAR detect no true positives (i.e. $TP \approx 0$). However, we do note that when using AIC for model selection, LVMA successfully detected some of the mediators.

We describe the properties of the procedures when the biomarkers are not normally distributed in Web Figures 13 - 18. We note that violations of normality did not affect the results. The power and relative performance of LVMA was essentially unchanged.

We describe the properties of the procedures when assuming the number of factors is 20 in web Figures

11 and 12. The performance of all procedures are robust to this misspecification of q.

We describe the properties of the procedures when using alternative criteria for model selection in Web Figures 5-10. The relative performance of the procedures is the same for all criteria (EBIC, AIC, or BIC), but the overall performance tends to be higher when using EBIC.

Finally, we use these simulations to highlight a few other observations. Web Figures 8-10 show that as the effect size increases and the number of TP approaches the maximum of 20, the bias of LVMA in estimating the conditional effect shrinks towards 0. Web Figures 19 - 21 show that the relative performance of LVMA in detecting mediators increases when the loadings of the exposure related factors decrease.

Web Appendix G Further Discussion of the Breast Cancer Study

Recall, in the main paper, LVMA identified only a single factor associated with both BMI ($\hat{\beta}_{EF,1} = 0.0045$) and risk of breast cancer ($\hat{\beta}_{FY,1} = 0.23$). This factor had 137 non-zero loadings but only 16 of these loadings had an absolute value larger than 0.4, with the majority of the remaining metabolites having loadings below 0.01. In Table 1, we list these 16 metabolites. In Web Figure 22, we display loadings for all metabolites from LVMA.

Here, we consider the results when using the other three methods: TMAO, TMAR and IMA. We note that neither TMAO or TMAR identified a statistically significant mediating factor and IMA did not identify statistically significant mediation metabolites. First, we consider TMAO. Web Figure 23 summarizes results for TMAO method with $q = 40$ factors. All p-values from individually testing each estimated factors were above 0.05. The most significant p-values (< 0.1) were for factors 1, 6, 9 and 17, (see Web Figure 23a) with, respectively 217, 304, 201, and 178 loadings having a magnitude (i.e. absolute value) greater than 0 (see Web Figure 23b) and 62, 31, 26 and 14 loadings having a magnitude larger than 0.2 (see Web Figure 23c). The next most significant factor (factor 8; p = 0.13) most closely resembled the factor identified by LVMA (see Web Figure 23d). Web Figure 24 summarizes results for TMAR method with $q = 40$ fitted factors. Second, we consider TMAR. The results were similar to those obtained from TMAO, with six factors having p-values below 0.1 (see Figure 24a). Again, factor 8 closely resembled the factor discovered by LVMA (see Web Figure 24d) and had a p-value of approximately 0.08.

Third, we consider IMA. IMA detected eight metabolites with p-value below 0.01 but none of them below Bonferroni threshold $0.05/481 = 0.0001$ (see Web Figure 25 and Web Table 1). Importantly, these findings reinforce our idea that joint modeling of the latent factors by LVMA can significantly increase the number of discovered mediators.

Web Appendix H Web Figures and Tables

Here, we present following additional figures and tables:

Web Figure 1: Causal graphs of the simulation models.

Web Figure 2: Continuous/Binary Outcome, EBIC selection and the exposure directly affects individual biomarker.

Web Figure 3: Continuous/Binary Outcome, BIC selection and the exposure directly affects individual biomarker.

Web Figure 4: Continuous/Binary Outcome, AIC selection and the exposure directly affects individual biomarker.

Web Figure 5: Results for a simulation model with continuous outcome, large factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. BIC criteria was used.

Web Figure 6: Results for a simulation model with binary outcome, large factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. BIC criteria was used.

Web Figure 7: Results for a simulation model with continuous and binary outcome, small factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. BIC criteria was used.

Web Figure 8: Results for a simulation model with continuous outcome, large factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. AIC criteria was used.

Web Figure 9: Results for a simulation model with binary outcome, large factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. AIC criteria was used.

Web Figure 10: Results for a simulation model with continuous and binary outcome, small factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. AIC criteria was used.

Web Figure 11: Results for a simulation model with continuous outcome, large factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. Model with $q = 20$ factors was fitted and EBIC criteria was used.

Web Figure 12: Results for a simulation model with binary outcome, large factor effects ($\lambda_1 = 0.3$), $N = 300$ and $N = 500$. Model with $q = 20$ factors was fitted and EBIC criteria was used.

Web Figure 13: Results for a simulation model with continuous outcome, non-symmetric and heavy tailed error. EBIC criteria was used.

Web Figure 14: Results for a simulation model with binary outcome, non-symmetric and heavy tailed error. BIC criteria was used.

Web Figure 15: Results for a simulation model with continuous outcome, non-symmetric and heavy tailed error. AIC criteria was used.

Web Figure 16: Results for a simulation model with binary outcome, left truncated error and $N = 500$. EBIC criteria was used.

Web Figure 17: Results for a simulation model with continuous and binary outcomes, left truncated error and $N = 500$. BIC criteria was used.

Web Figure 18: Results for a simulation model with continuous and binary outcomes, left truncated error and $N = 500$. AIC criteria was used.

Web Figure 19: Results for a simulation model with continuous outcome, large factor effects ($\lambda_1 = 0.3$), small exposure effects on factors ($\beta_{EF,2} = 0.25$), $N = 300$ and $N = 500$. EBIC criteria was used.

Web Figure 20: Results for a simulation model with binary outcome, large factor effects ($\lambda_1 = 0.3$), small exposure effects on factors ($\beta_{EF,2} = 0.25$), $N = 300$ and $N = 500$. EBIC criteria was used.

Web Figure 21: Results for a simulation model with continuous and binary outcomes, small factor effects ($\lambda_1 = 0.25$), small exposure effects on factors ($\beta_{EF,2} = 0.25$) and $N = 500$. EBIC criteria was used.

Web Figure 22: Loading affects on metabolites by the factor mediating increased BMI and ER+ breast cancer in PLCO.

Web Figure 23: Summary results of TMAO method for detecting factors that mediate increased BMI and ER+ breast cancer in PLCO.

Web Figure 24: Summary results of TMAR method for detecting factors that mediate increased BMI and ER+ breast cancer in PLCO.

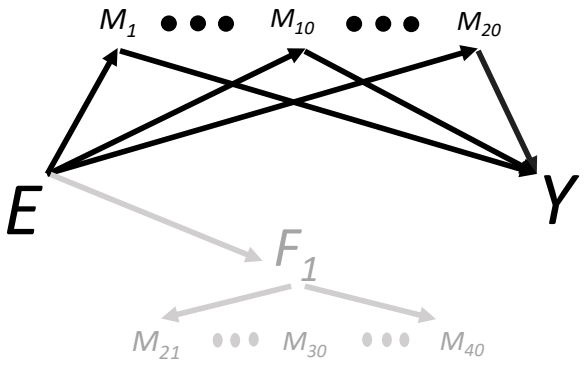
Web Figure 25: Summary results of IMA method for detecting metabolites that mediate increased BMI and ER+ breast cancer in PLCO.

Web Table 1: List of metabolites with p-values below 0.01.

Web Table 2: Comparison of computational time of TMAO, TMAR and LVMA.

H.1 Figures

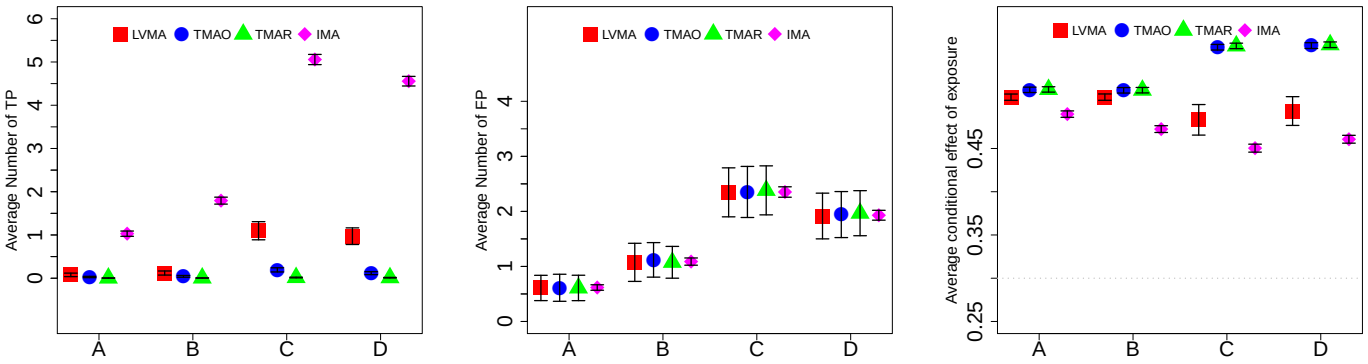
Web Figure 1: **Causal graph of the simulation models.** Our model assumes that only first set of 20 biomarkers directly mediate the effect of E on Y. The first four latent factors are only associated with E and the remaining ten are associated neither with E nor Y.



Web Figure 2: **Continuous/Binary Outcome, EBIC selection and the exposure directly affects individual biomarker**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EM,1} = 0.12$, $\lambda_0 = 0.4$; B: $\beta_{EM,1} = 0.12$, $\lambda_0 = 0.5$; C: $\beta_{EM,1} = 0.15$, $\lambda_0 = 0.4$; D: $\beta_{EM,1} = 0.15$, $\lambda_0 = 0.5$) based on 1000 simulations. Te top panel is for a study with a continuous outcome with $N = 500$ and the bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

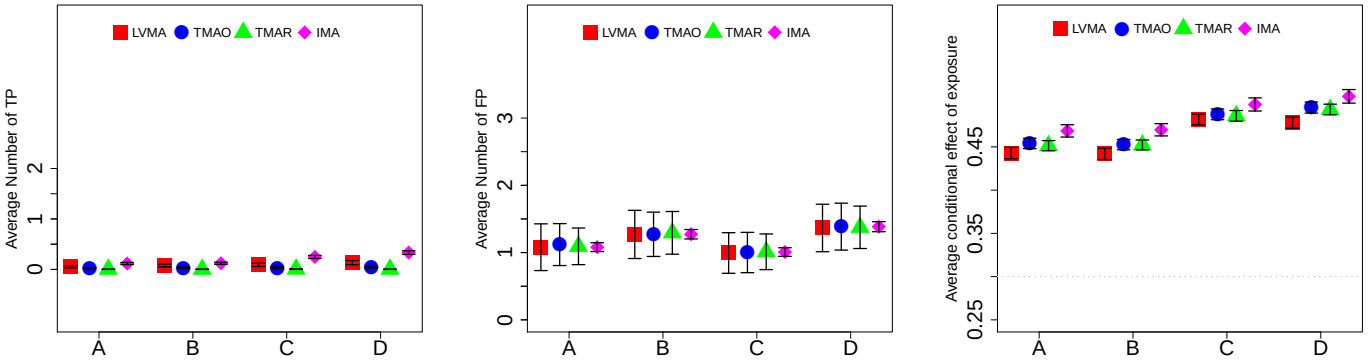
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), EBIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), EBIC selection

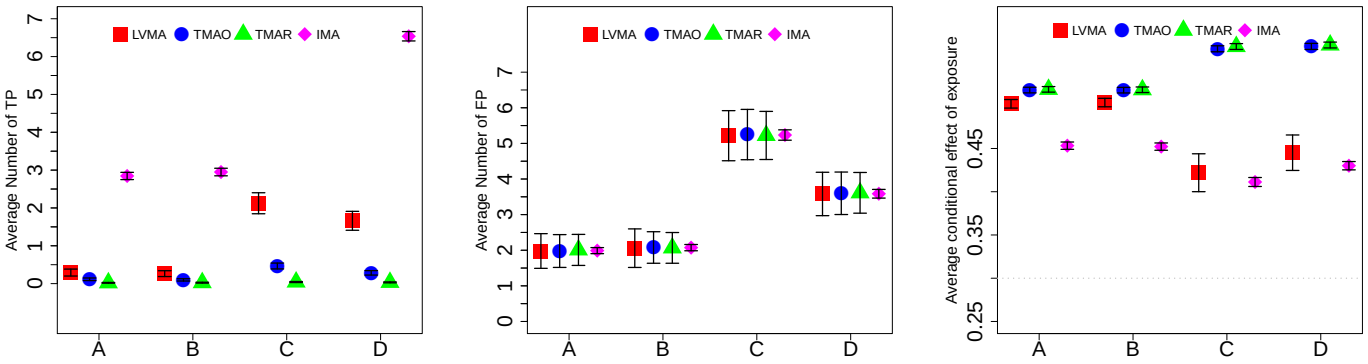
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 3: **Continuous/Binary Outcome, BIC selection and the exposure directly affects individual biomarker**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EM,1} = 0.12$, $\lambda_0 = 0.4$; B: $\beta_{EM,1} = 0.12$, $\lambda_0 = 0.5$; C: $\beta_{EM,1} = 0.15$, $\lambda_0 = 0.4$; D: $\beta_{EM,1} = 0.15$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

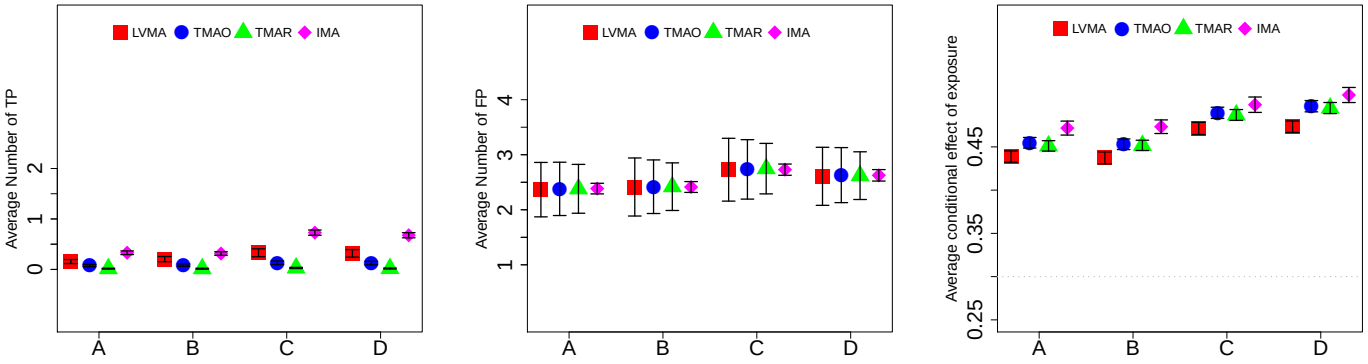
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), BIC selection

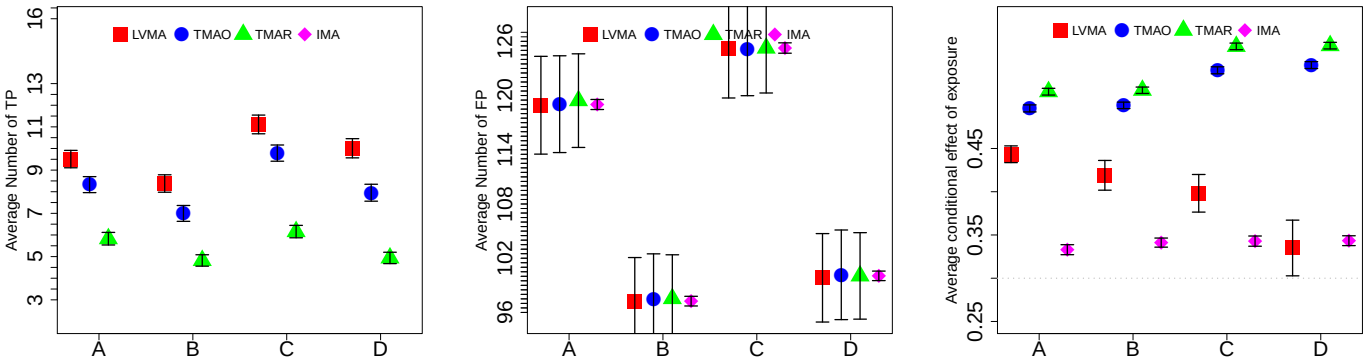
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 4: **Continuous/Binary Outcome, AIC selection and the exposure directly affects individual biomarker**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EM,1} = 0.12$, $\lambda_0 = 0.4$; B: $\beta_{EM,1} = 0.12$, $\lambda_0 = 0.5$; C: $\beta_{EM,1} = 0.15$, $\lambda_0 = 0.4$; D: $\beta_{EM,1} = 0.15$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

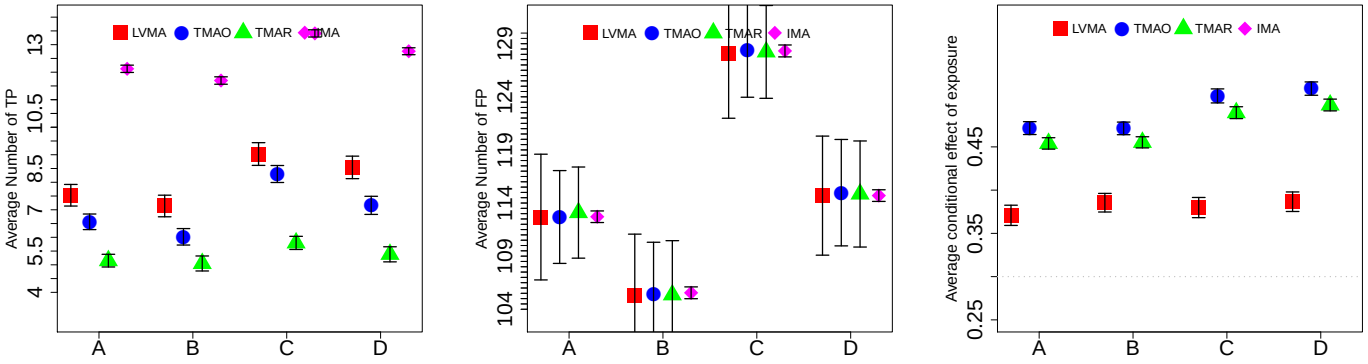
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), AIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), AIC selection

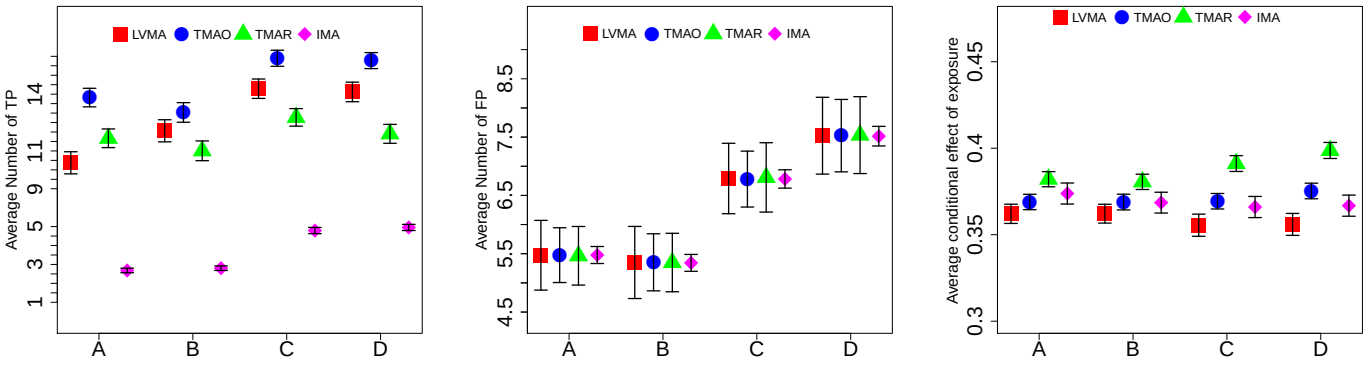
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 5: **Continuous Outcome and Large Factor Effects ($\lambda_1 = 0.3$), BIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

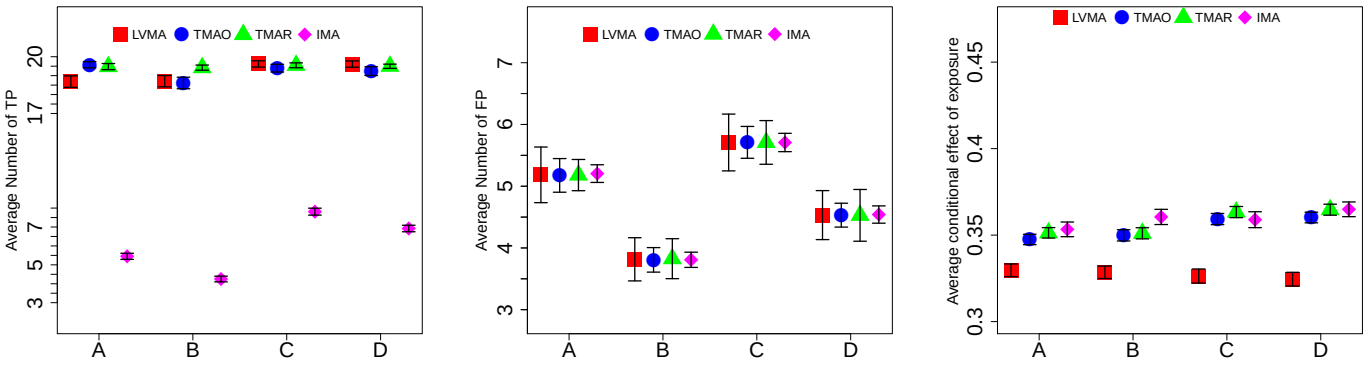
Continuous outcome, $N = 300$, BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Continuous outcome, $N = 500$, BIC selection

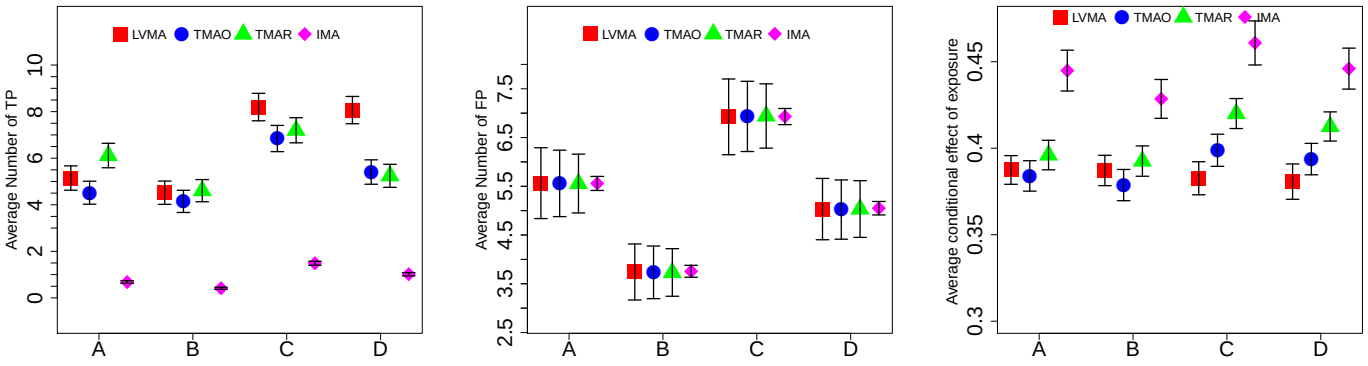
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 6: **Binary Outcome and Large Factor Effects ($\lambda_1 = 0.3$), BIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

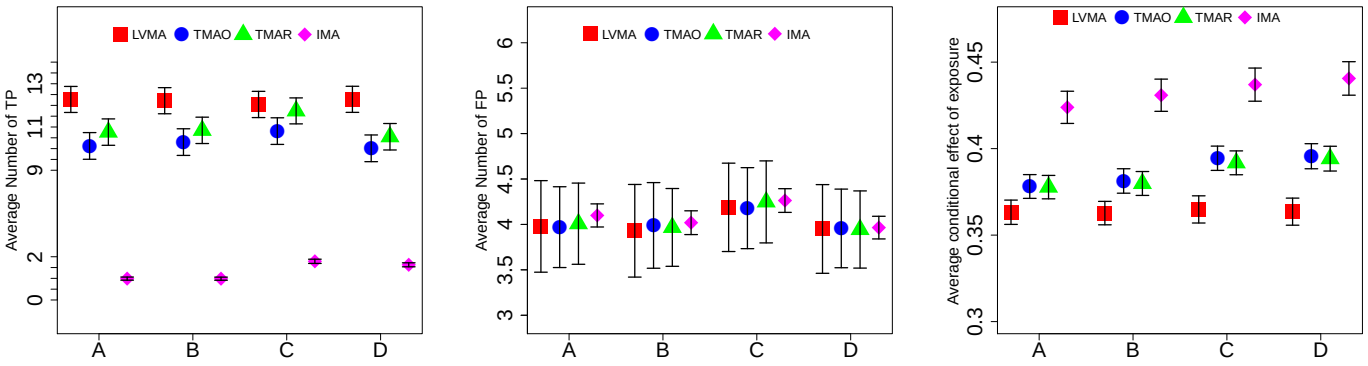
Binary outcome, $N_1 = N_0 = 150$, BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, BIC selection

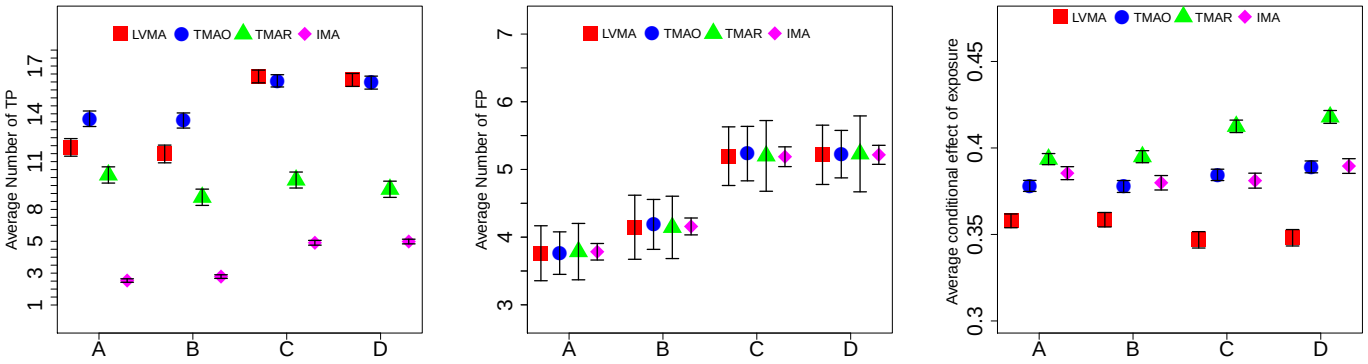
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 7: **Small Factor Effects ($\lambda_1 = 0.25$), BIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with continuous and binary outcomes, respectively. The whiskers show two standard errors around the average estimates.

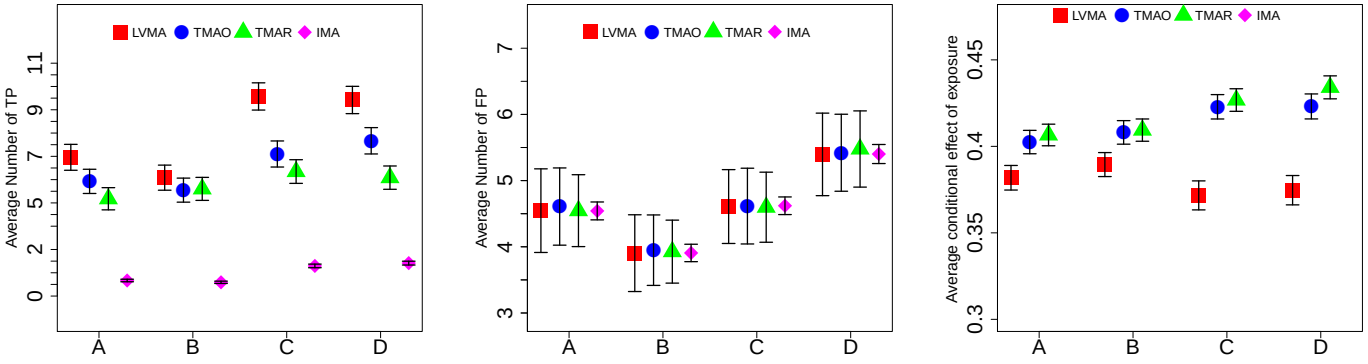
Continuous outcome, $N = 500$, BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, BIC selection

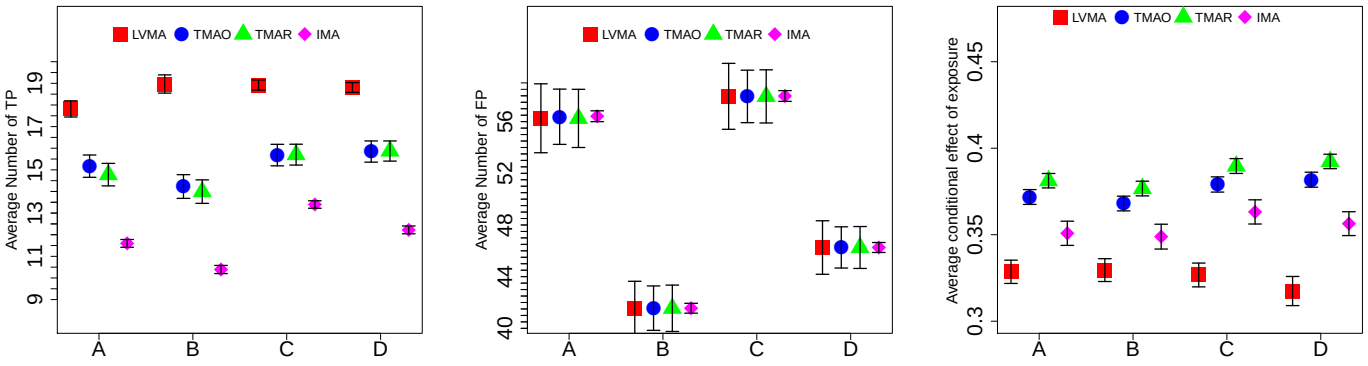
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 8: **Continuous Outcome and Large Factor Effects ($\lambda_1 = 0.3$), AIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

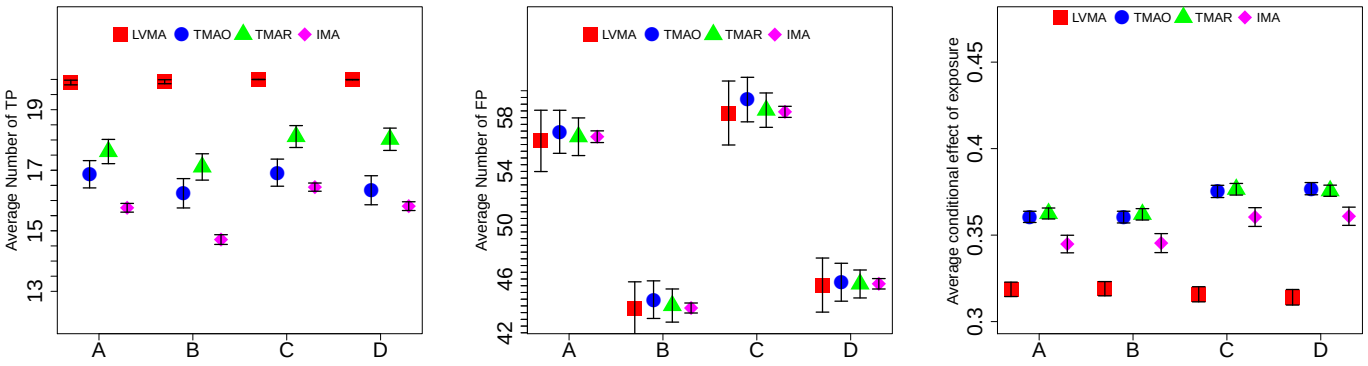
Continuous outcome, $N = 300$, AIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Continuous outcome, $N = 500$, AIC selection

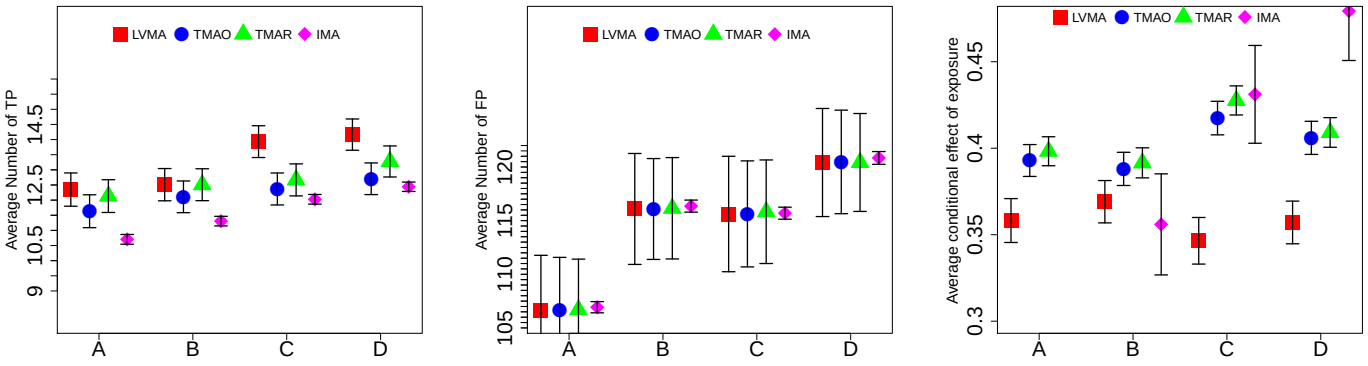
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 9: **Binary Outcome and Large Factor Effects ($\lambda_1 = 0.3$), AIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

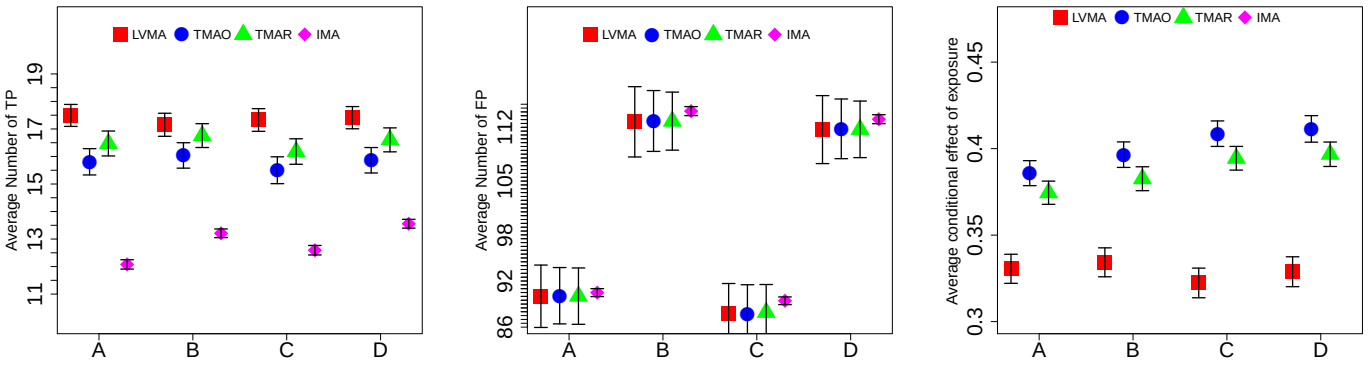
Binary outcome, $N_1 = N_0 = 150$, AIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, AIC selection

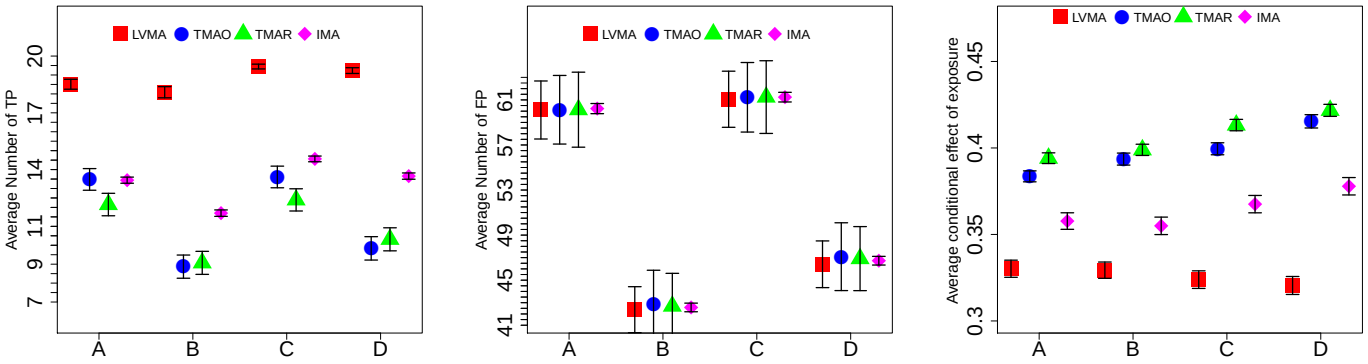
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 10: **Small Factor Effects ($\lambda_1 = 0.25$), AIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with continuous and binary outcomes, respectively. The whiskers show two standard errors around the average estimates.

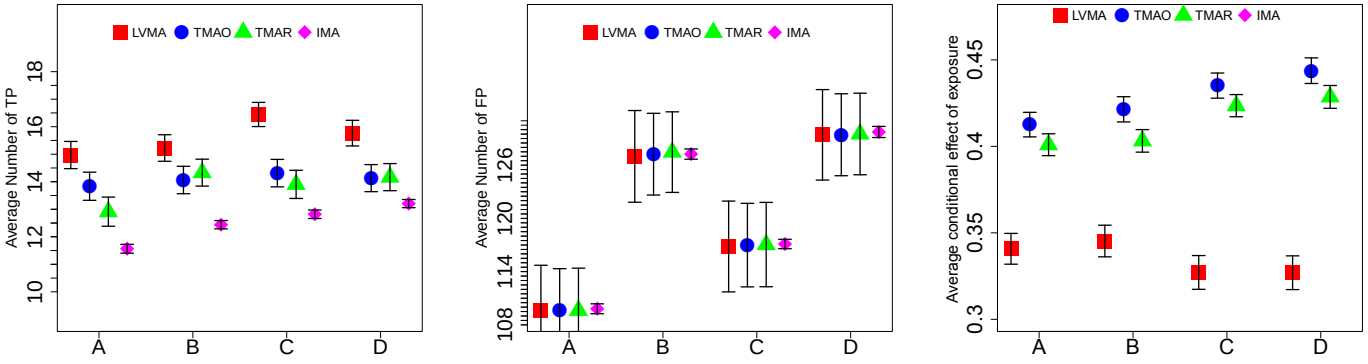
Continuous outcome, $N = 500$, AIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, AIC selection

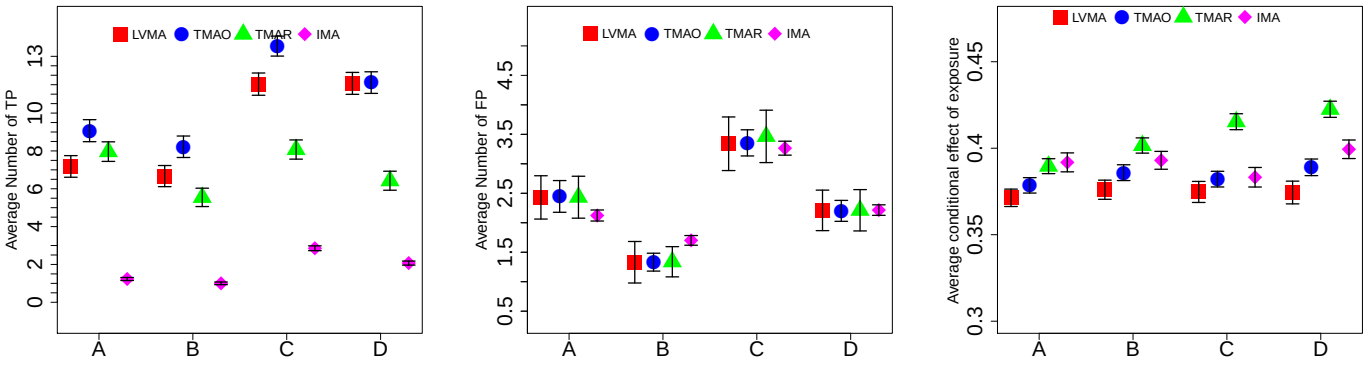
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 11: **Continuous Outcome and Large Factor Effects ($\lambda_1 = 0.3$), EBIC selection and fitted $q = 20$ factors**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

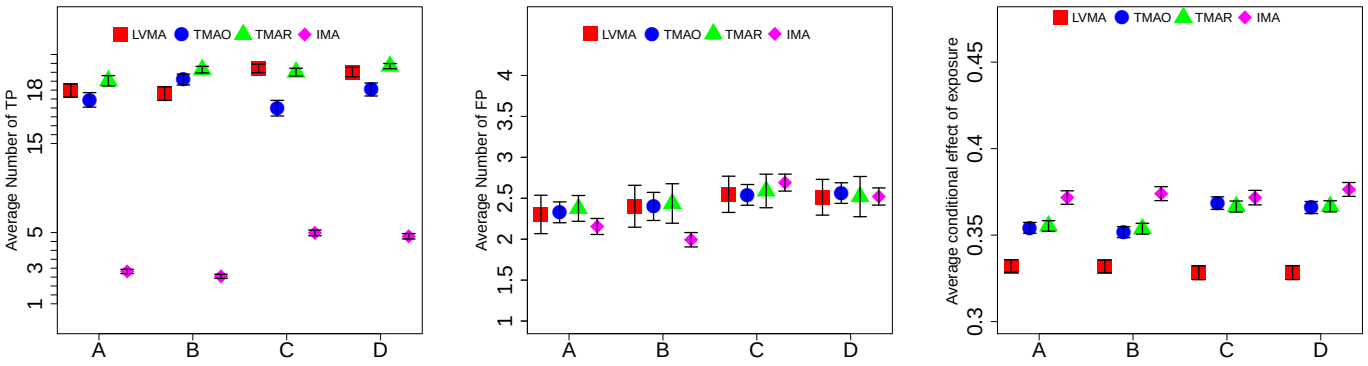
Continuous outcome, $N = 300$, EBIC selection, $q = 20$

a) True Positive b) False Positive c) Estimate of Conditional Effect



Continuous outcome, $N = 500$, EBIC selection, $q = 20$

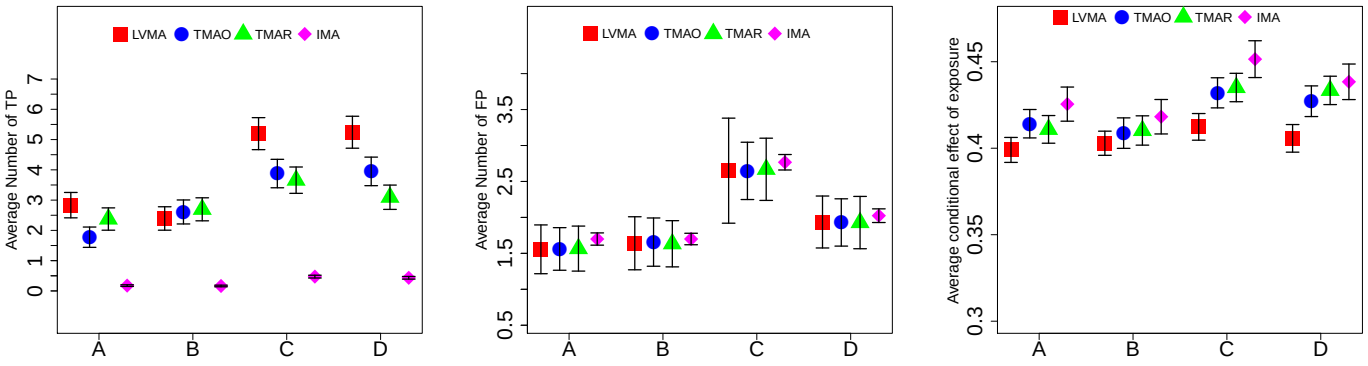
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 12: **Binary Outcome and Large Factor Effects ($\lambda_1 = 0.3$), EBIC selection and fitted $q = 20$ factors**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

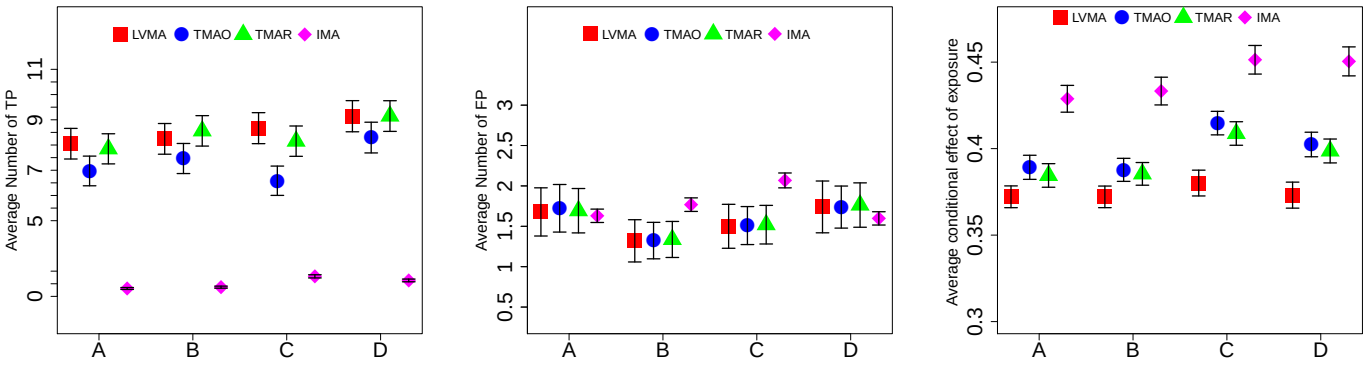
Binary outcome, $N_1 = N_0 = 150$, EBIC selection, $q = 20$

a) True Positive b) False Positive c) Estimate of Conditional Effect

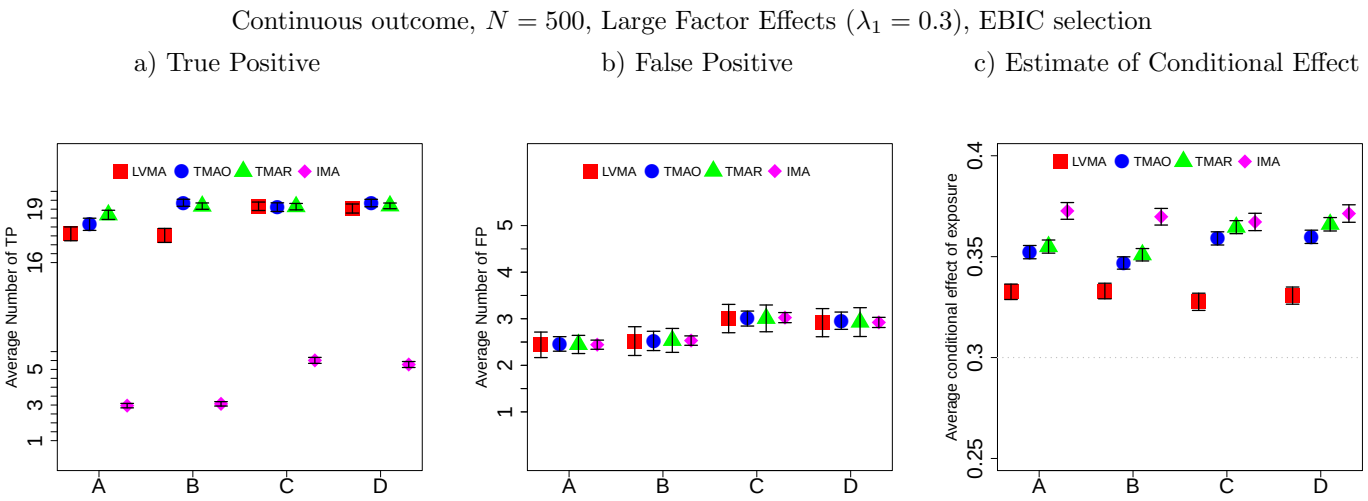


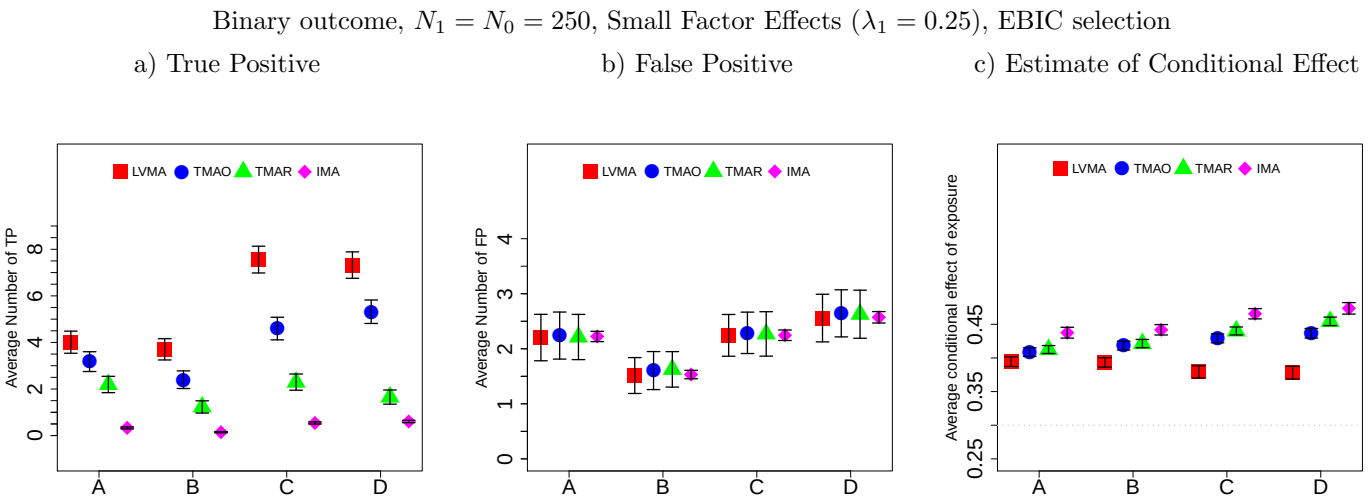
Binary outcome, $N_1 = N_0 = 250$, EBIC selection, $q = 20$

a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 13: **Continuous/Binary Outcome, non-symmetric and heavy tailed error in biomarker, EBIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

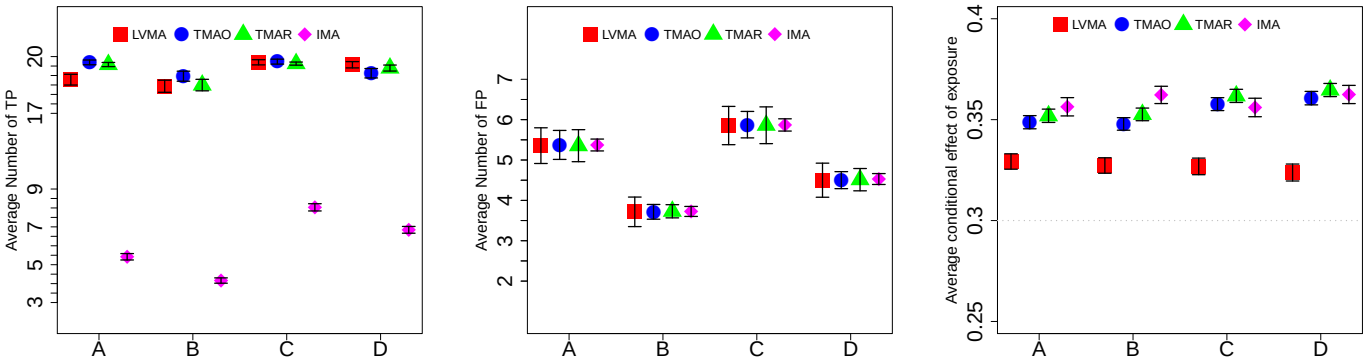




Web Figure 14: **Continuous/Binary Outcome, non-symmetric and heavy tailed error in biomarker, BIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

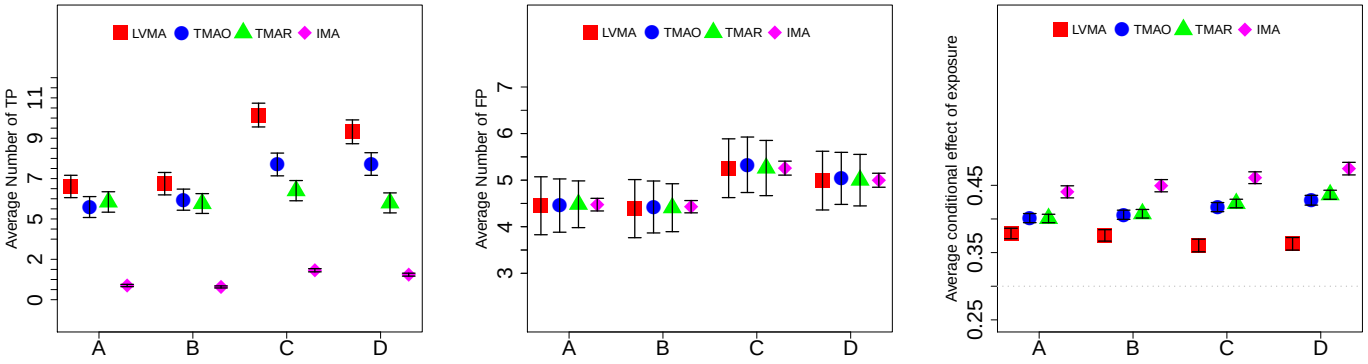
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect

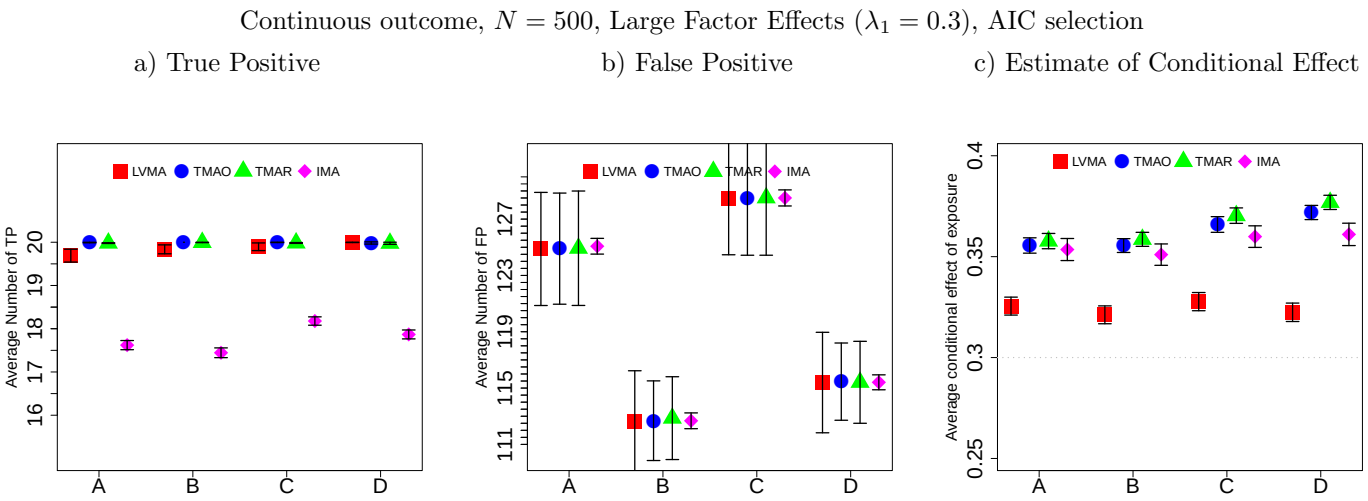


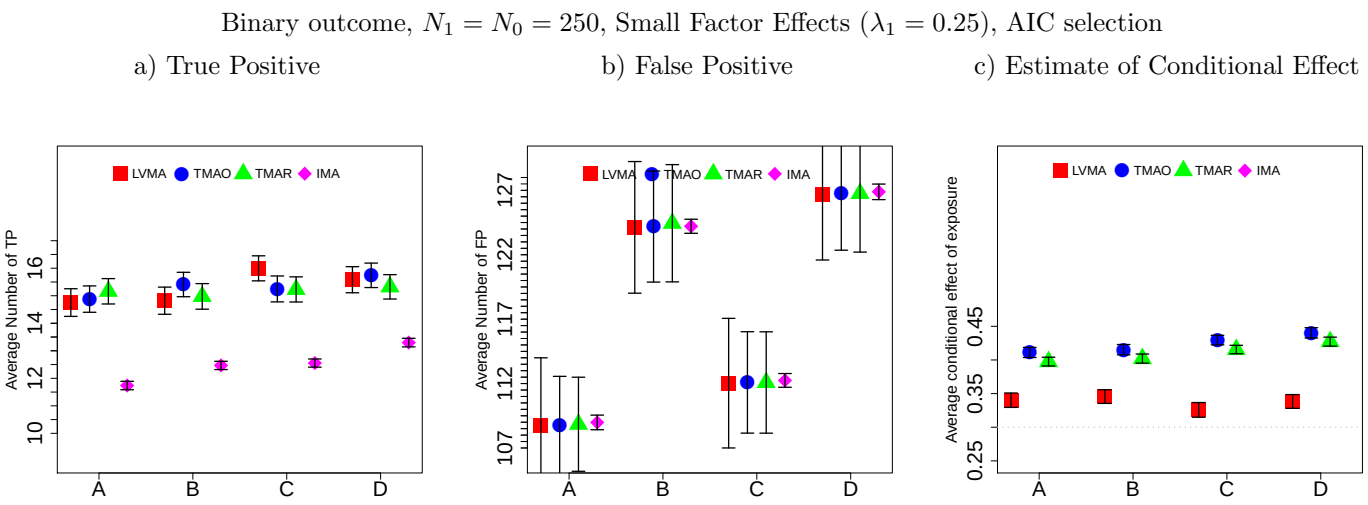
Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 15: **Continuous/Binary Outcome, non-symmetric and heavy tailed error in biomarker, AIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

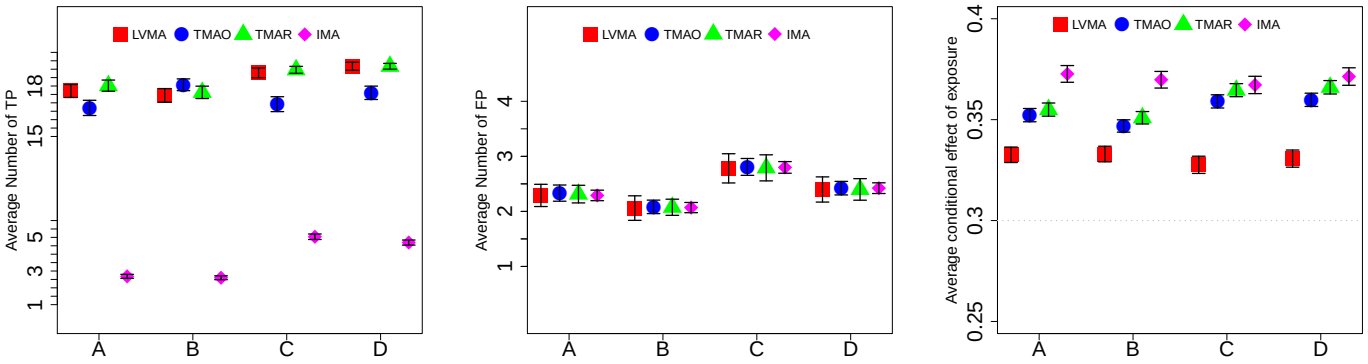




Web Figure 16: **Continuous/Binary Outcome and left truncated error in biomarker, EBIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

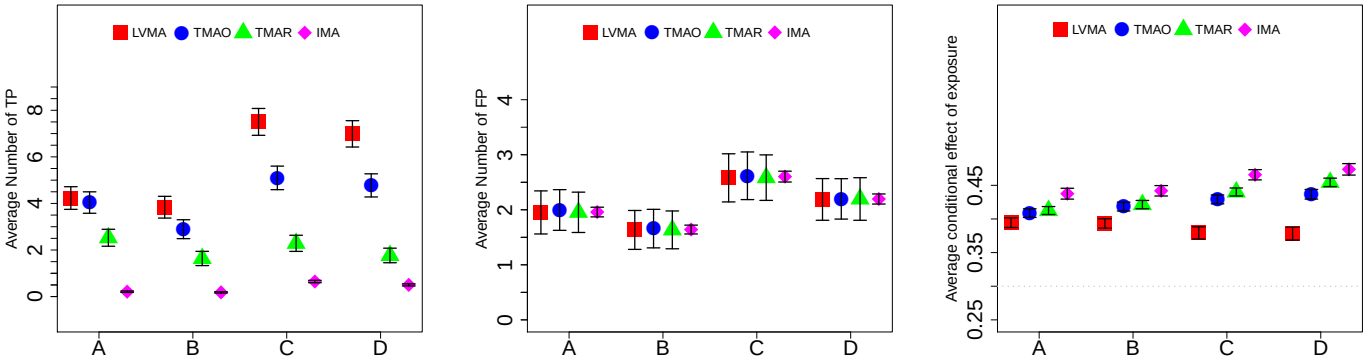
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), EBIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), EBIC selection

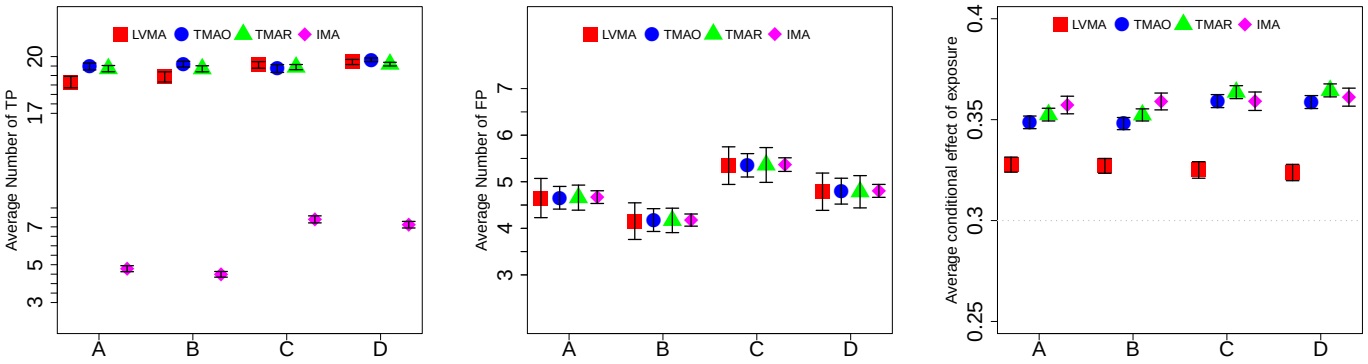
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 17: **Continuous/Binary Outcome and left truncated error in biomarker, BIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

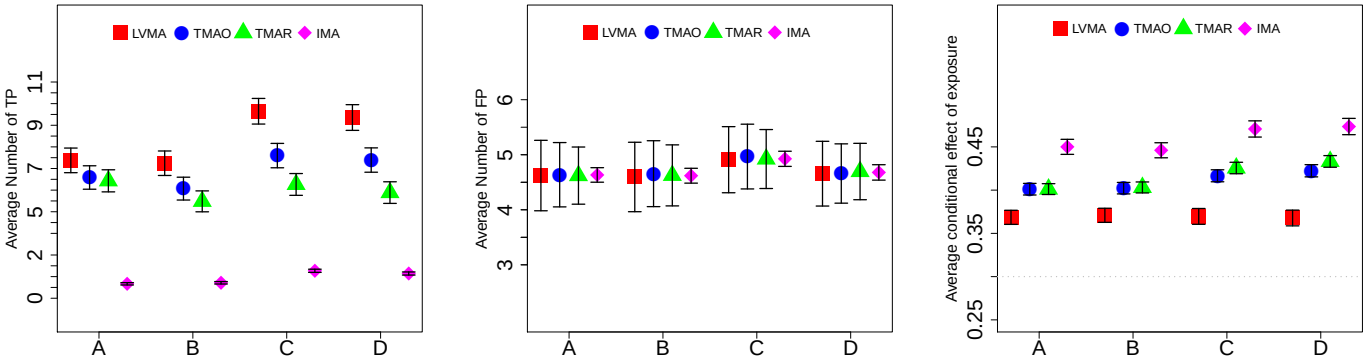
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), BIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), BIC selection

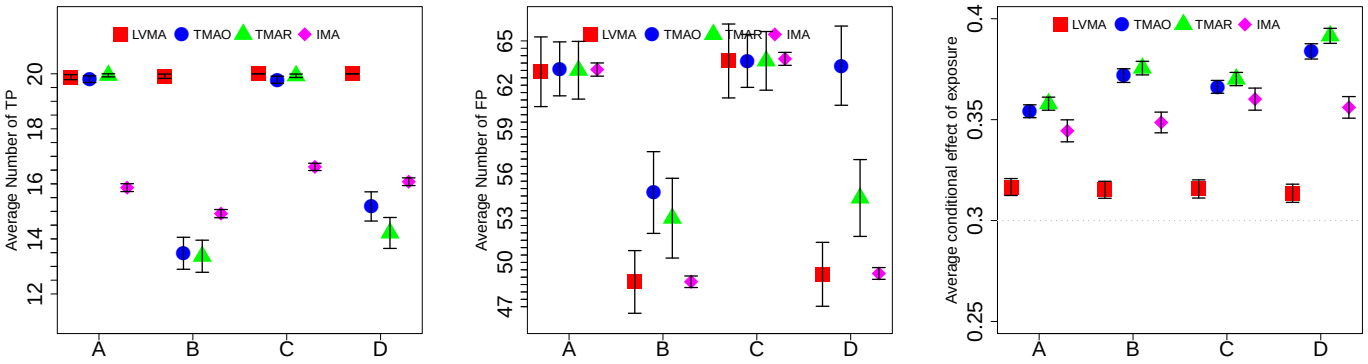
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 18: **Continuous/Binary Outcome and left truncated error in biomarker, AIC selection**. The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top panel is for a study with continuous outcome with $N = 500$ and bottom panel is for a study with binary outcome with $N_1 = N_0 = 250$ subjects respectively. The whiskers show two standard errors around the average estimates.

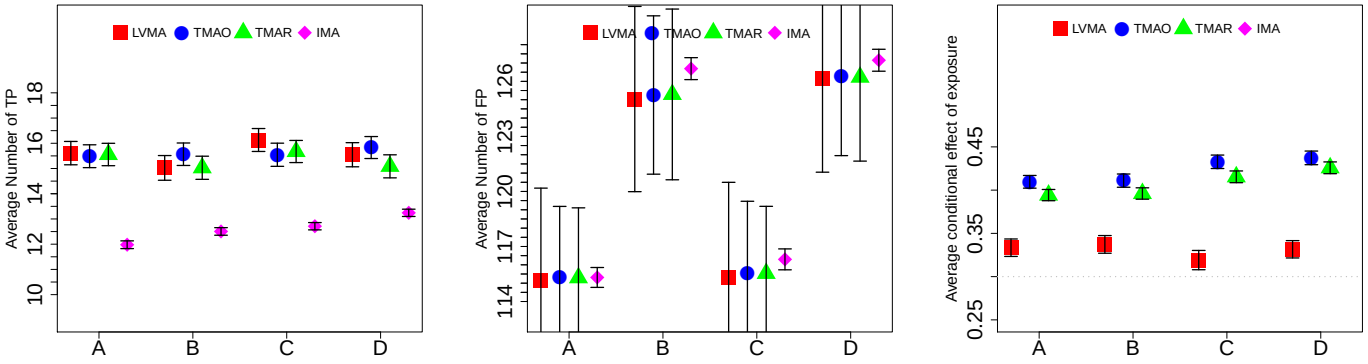
Continuous outcome, $N = 500$, Large Factor Effects ($\lambda_1 = 0.3$), AIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, Small Factor Effects ($\lambda_1 = 0.25$), AIC selection

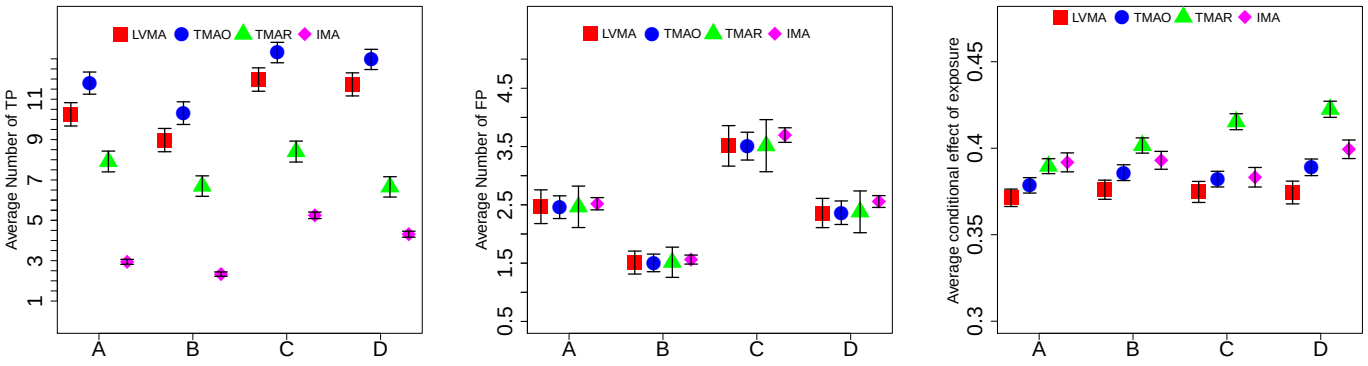
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 19: **Continuous Outcome and Large Factor Effects ($\lambda_1 = 0.3$), EBIC selection and Small Exposure Effects ($\beta_{EF,2} = 0.4$).** The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

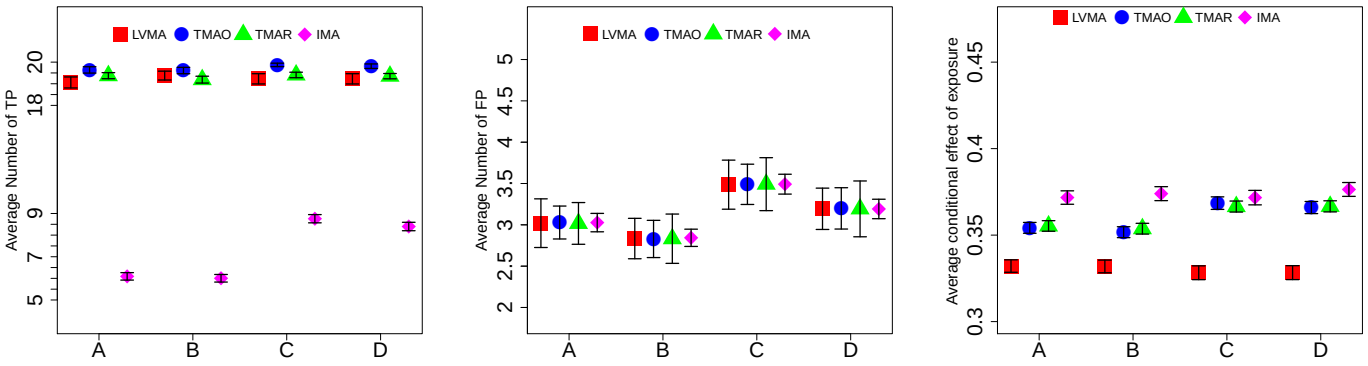
Continuous outcome, $N = 300$, EBIC selection

a) True Positive b) False Positive c) Average Estimate Direct Effect



Continuous outcome, $N = 500$, EBIC selection

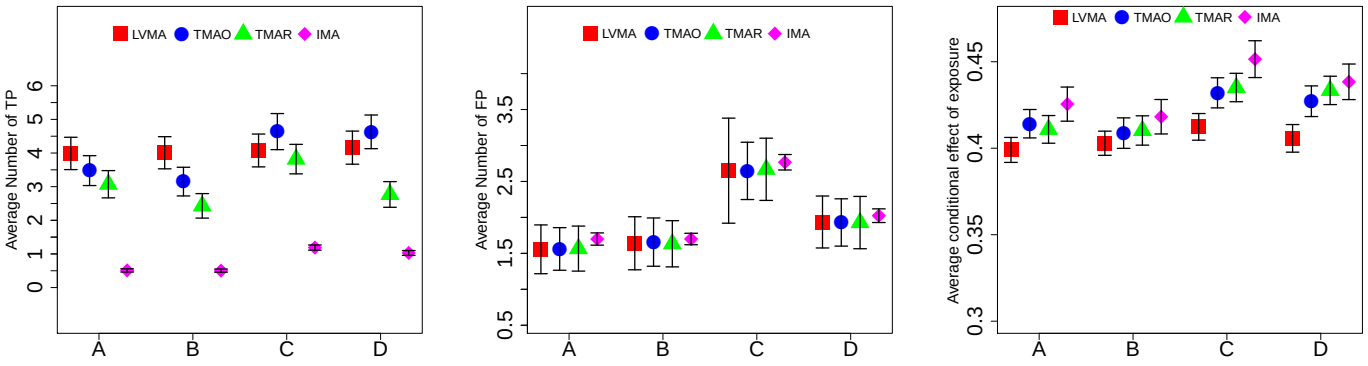
a) True Positive b) False Positive c) Average Estimate Direct Effect



Web Figure 20: **Binary Outcome and Large Factor Effects ($\lambda_1 = 0.3$), EBIC selection and Small Exposure Effects ($\beta_{EF,2} = 0.4$).** The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with N = 300 and N = 500 subjects respectively. The whiskers show two standard errors around the average estimates.

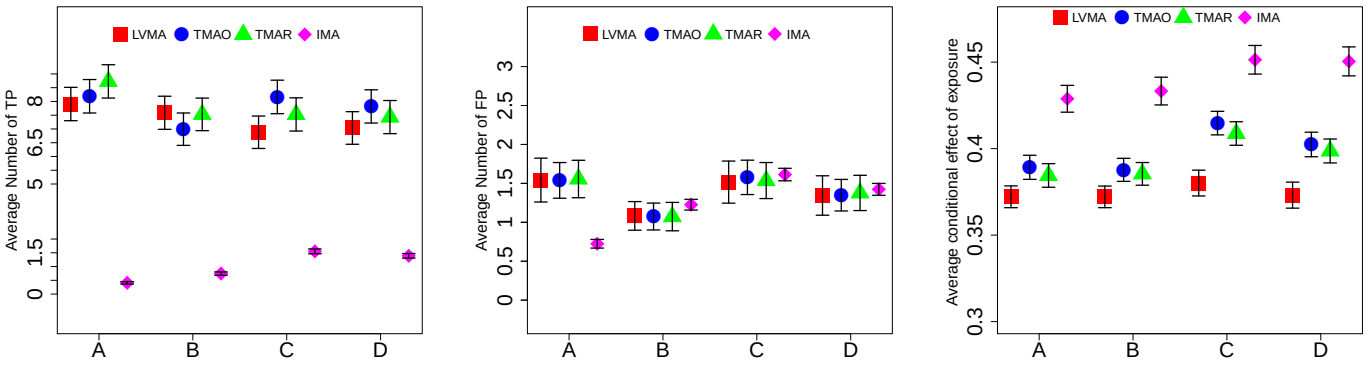
Binary outcome, $N_1 = N_0 = 150$, EBIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



Binary outcome, $N_1 = N_0 = 250$, EBIC selection

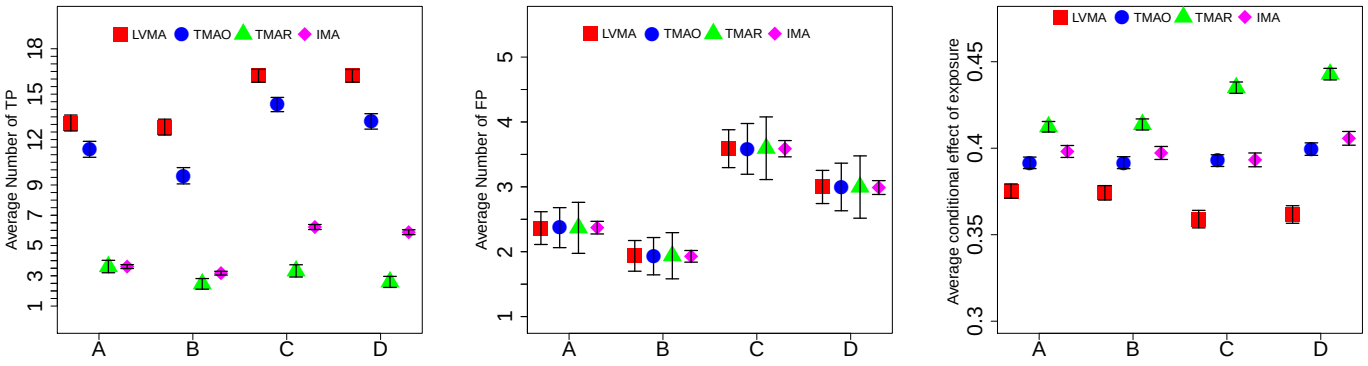
a) True Positive b) False Positive c) Estimate of Conditional Effect



Web Figure 21: **Small Factor Effects ($\lambda_1 = 0.25$), EBIC selection and Small Exposure Effects ($\beta_{EF,2} = 0.4$).** The panels, labeled a-c, show the average number of true positives (TP), the average number of false positives (FP), and the average estimate of the direct effect for the four methods (red=LVMA; blue=TMAO; green=TMAR; purple=IMA) and for four scenarios (A: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.4$; B: $\beta_{EF,1} = 0.4$, $\lambda_0 = 0.5$; C: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.4$; D: $\beta_{EF,1} = 0.5$, $\lambda_0 = 0.5$) based on 1000 simulations. Top and bottom panel are for studies with continuous and binary outcomes, respectively ($N = 500$). The whiskers show two standard errors around the average estimates.

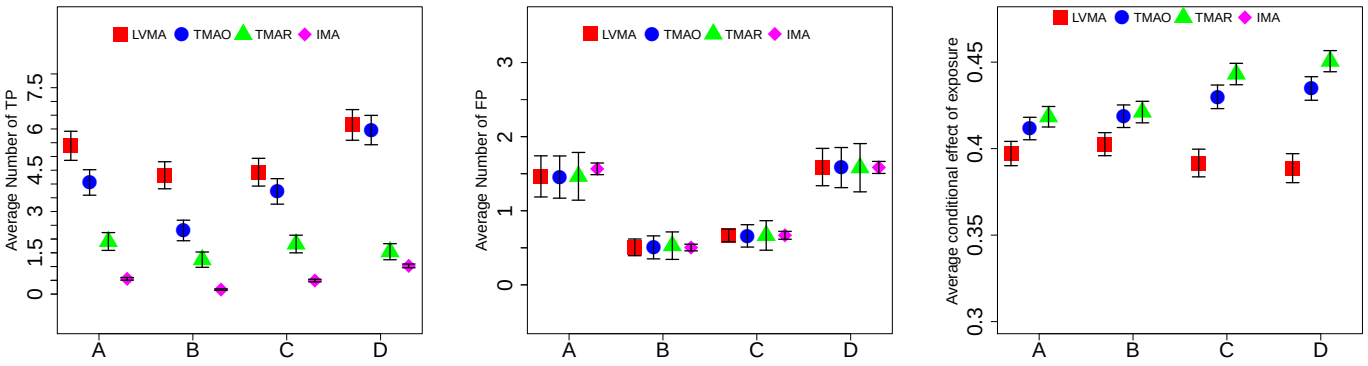
Continuous outcome, $N = 500$, EBIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect



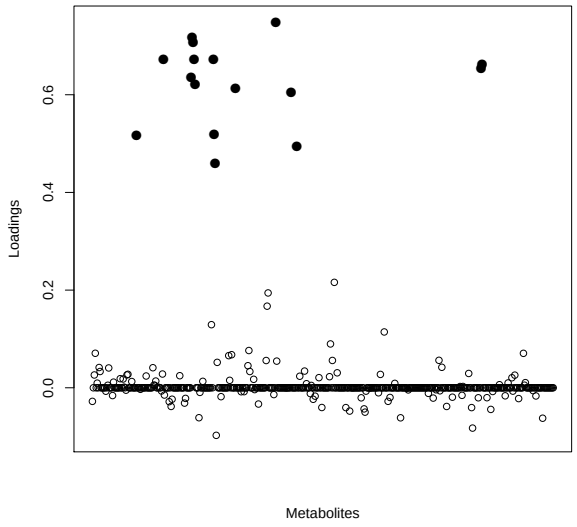
Binary outcome, $N_1 = N_0 = 250$, EBIC selection

a) True Positive b) False Positive c) Estimate of Conditional Effect

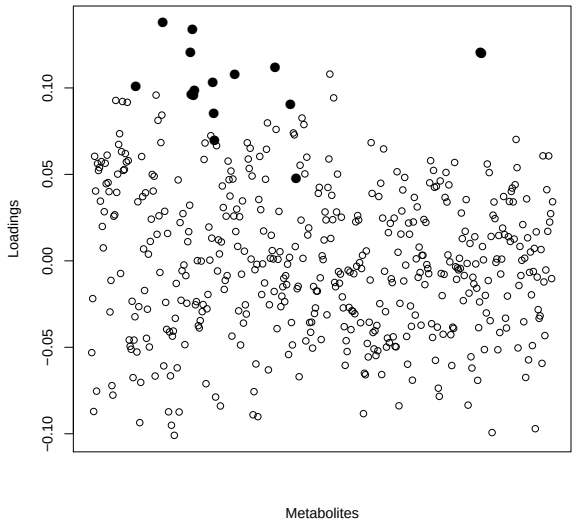


Web Figure 22: **Metabolites Linking BMI and Breast Cancer.** These figures include affects on metabolites by the factor mediating increased BMI and ER+ breast cancer in PLCO.

a) LVMA

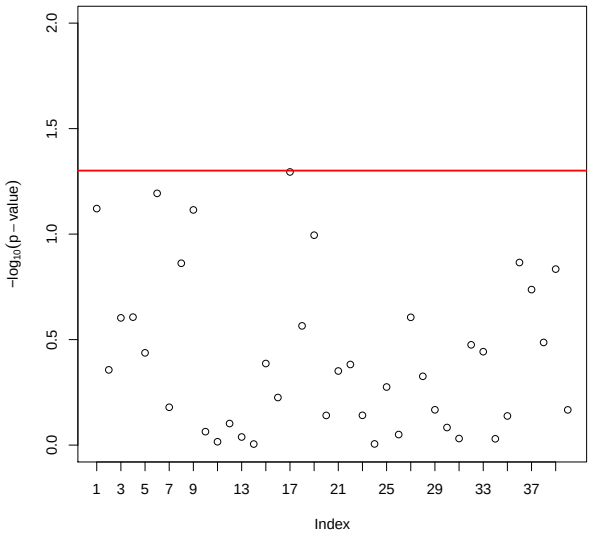


b) Standard Factor Analysis

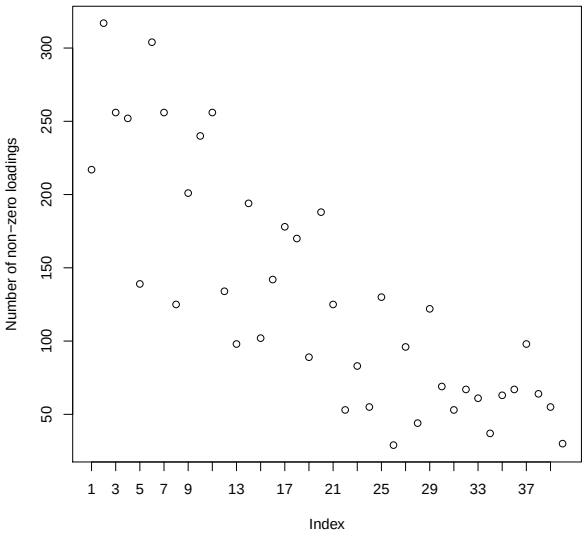


Web Figure 23: **Summary for TMAO**. Summary results for TMAO method with $q_{max} = 40$ fitted factor. P-values are calculated using Sobel's test on estimated factors.

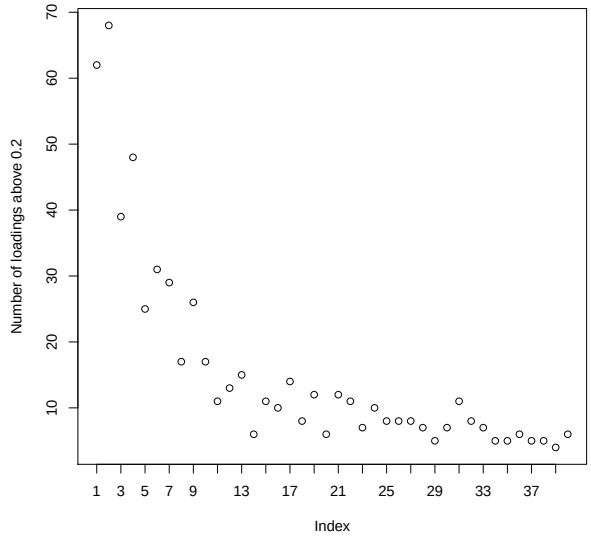
a) $-log_{10}(p - values)$



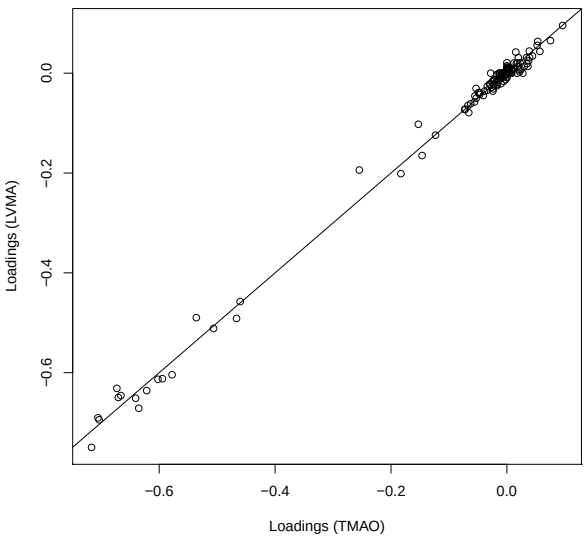
b) Numeber of non-zero loadings per factor



c) Number of loadings per factor larger than 0.2

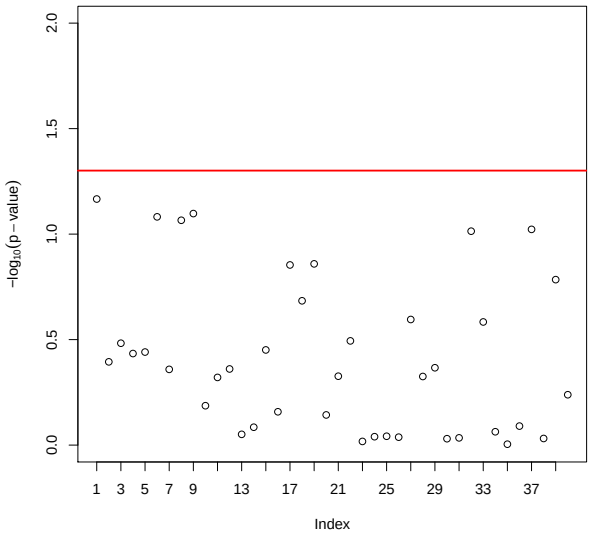


d) Comparison of factor loadings

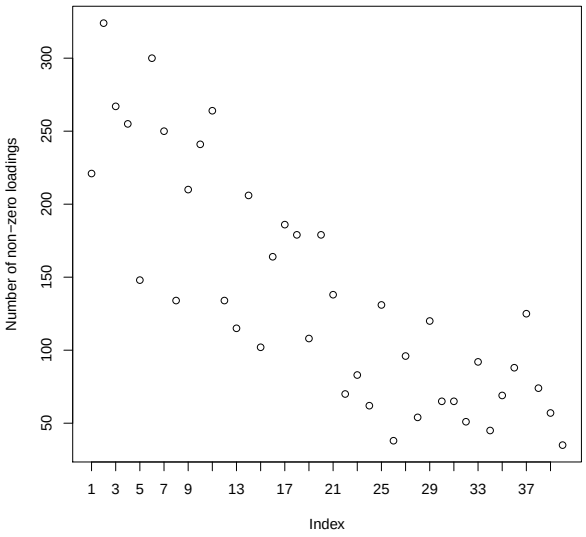


Web Figure 24: **Summary for TMAR**. Summary results for TMAR method with $q_{max} = 40$ fitted factor. P-values are calculated using Sobel's test on estimated factors.

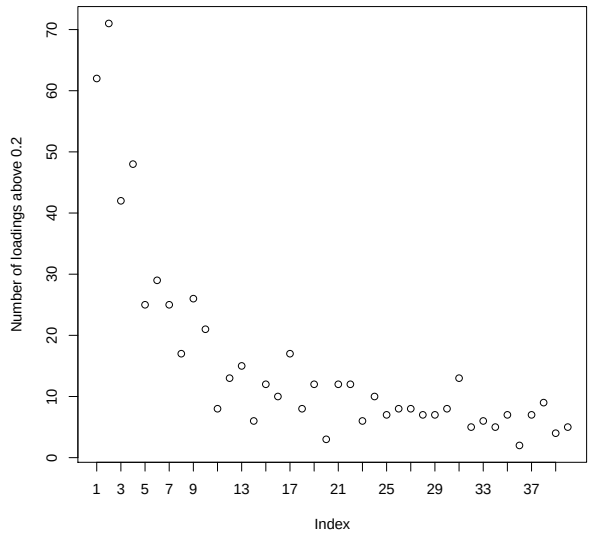
a) $-log_{10}(p - values)$



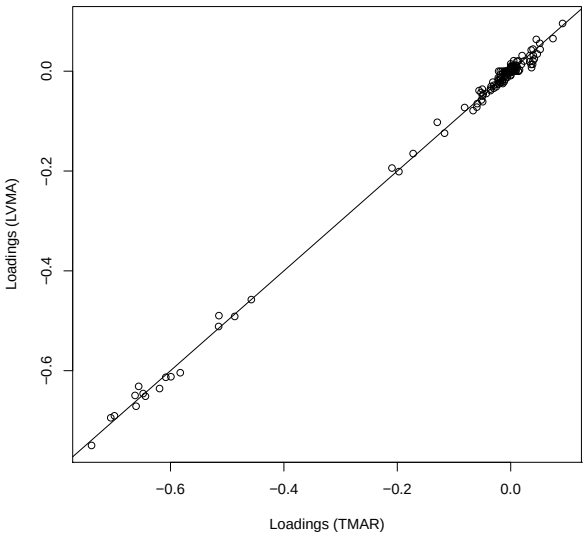
b) Numeber of non-zero loadings per factor



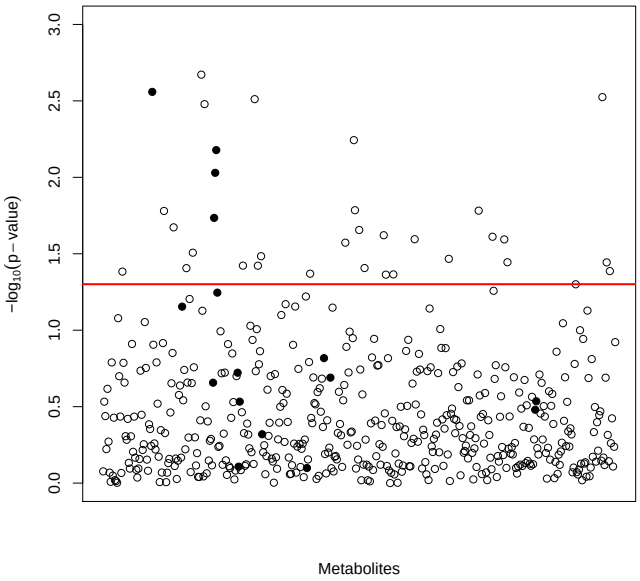
c) Number of loadings per factor larger than 0.2



d) Comparison of factor loadings



Web Figure 25: **Summary for IMA**. IMA's p-values are calculated using Sobel's test on each metabolite.



Web Table 1: **Metabolites Linking BMI and Breast Cancer.** This list includes those metabolites with p-values from Sobel's test below 0.01.

Metabolite	Loading (λ_{ij})
16α-hydroxy DHEA 3-sulfate	0.003
3-methyl-2-oxobutyrate	0.002
3- methylglutarylcarnitine	0.003
4-androsten-3β, 17-β-diol disulfate (2)	0.009
4-androsten-3β, 17-β-diol disulfate (1)	0.007
Allo-isoleucine	0.003
γ-glutamylvaline	0.006
Urate	0.003

Web Table 2: **Comparison of computational time.**

Number of Factors	TMAO	TMAR	LVMA
15	24 min	24 min	10hr
20	33 min	32 min	12hr
40	1:12hr	1:10hr	26hr

References

[1] Albert, J. M., Geng, C., and Nelson, S. (2016). Causal mediation analysis with a latent mediator. *Biometrical Journal*, 58(3):535–548.

[2] Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150.

[3] Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180.

[4] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

[5] Hirose, K. (2016). Simple structure estimation via prenet penalization. *arXiv preprint arXiv:1607.01145*.

[6] Hirose, K. and Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, 79:120–132.

[7] Küchenhoff, H. (1995). The identification of logistic regression models with errors in the variables. *Statistical Papers*, 36(1):41–47.

[8] Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

[9] Pfeiffer, R. M. and Riedl, R. (2015). On the use and misuse of scalar scores of confounders in design and analysis of observational studies. *Statistics in Medicine*, 34(18):2618–2635.

[10] Rubin, D. B. and Thayer, D. T. (1982). Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76.

[11] Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678.

[12] Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017). Expandable factor analysis. *Biometrika*, 104(3):649–663.

[13] VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2016). Mediation analysis with matched case-control study designs. *American Journal of Epidemiology*, 183(9):869—870.

[14] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509.