Biostatistics (2016), **0**, 0, *pp.* 1–18
doi:10.1093/biostatistics/GuidedBAC-Supplementary

Guided Bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. Supplementary materials

JOSEPH ANTONELLI*, CORWIN ZIGLER, AND FRANCESCA DOMINICI

Department of Biostatistics, Harvard TH Chan School of Public Health 677 Huntington Avenue, Boston, MA, 02115

jantonel@hsph.harvard.edu

APPENDIX

A. Details of posterior calculation

Here we show the full posterior and corresponding conditionals for implementing a gibbs sampler when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are either binary or continuous. For simplifying notation we will ignore $\boldsymbol{C_{mis}}$ and let the matrix $\boldsymbol{C}$ represent all covariates in the data, while acknowledging that some of these covariates are missing in a subset of the subjects. First let $\boldsymbol{\theta^y} = [\theta_0^y, \beta, \beta^S, \theta_1^y, ..., \theta_p^y, \theta_1^{yS}, ..., \theta_M^{y\ S}]$ and $\boldsymbol{\theta^x} = [\theta_0^x, \theta_1^x, ..., \theta_p^x, \theta_1^{xS}, ..., \theta_M^{x\ S}]$. Then Letting X_i^*, Y_i^*, and C_{ij}^* be latent variables for X_i, Y_i, and C_{ij} respectively the posterior can be written as follows:

*To whom correspondence should be addressed.

$$
\begin{aligned}
&P(\boldsymbol{\theta^y}, \boldsymbol{\theta^x}, \boldsymbol{\theta^{M+1}}, ..., \boldsymbol{\theta^P}, \sigma_y^2, \sigma_x^2, \sigma_{M+1}^2, ..., \sigma_P^2, \boldsymbol{\alpha^x}, \boldsymbol{\alpha^y}, \boldsymbol{\alpha^{M+1}}, ..., \boldsymbol{\alpha^P}, \boldsymbol{C^*}, \boldsymbol{Y^*}, \boldsymbol{X^*} | \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{C}) \\
&\propto \prod_{i=1}^{N} p(Y_i|Y_i^*) p(Y_i^*|\boldsymbol{\theta^y}, X_i, \boldsymbol{C_i}, \sigma_y^2, \boldsymbol{\alpha^y}) \\
&\times\ p(X_i|X_i^*) p(X_i^*|\boldsymbol{\theta^x}, \boldsymbol{C_i}, \sigma_x^2, \boldsymbol{\alpha^x}) \\
&\times \prod_{j=M+1}^{P} p(C_{ij}|C_{ij}^*) p(C_{ij}^*|\boldsymbol{\theta^j}, \sigma_j^2, \boldsymbol{\alpha^j}, \boldsymbol{C_i^*}) \\
&\times P(\boldsymbol{\theta^y}) P(\boldsymbol{\theta^x}) P(\boldsymbol{\theta^j}) P(\sigma_y^2) P(\sigma_x^2) P(\boldsymbol{\sigma_j^2}) P(\boldsymbol{\alpha^y}, \boldsymbol{\alpha^x}) P(\boldsymbol{\alpha^j})
\end{aligned}
$$

$$
\begin{aligned}
&\propto \prod_{i=1}^{N} p(Y_i|Y_i^*) N(Y_i^*; \theta_0^y + \beta X_i + \beta^S X_i S_i + \sum_{p=1}^{P} \alpha_p^y \theta_p^y C_{ip} + \sum_{p=1}^{M} \alpha_p^y \theta_p^{yS} C_{ip} S_i, \sigma_y^2) \\
&\times \prod_{i=1}^{N} p(X_i|X_i^*) N(X_i^*; \theta_0^x + \sum_{p=1}^{P} \alpha_p^x \theta_p^x C_{ip} + \sum_{p=1}^{M} \alpha_p^x \theta_p^{xS} C_{ip} S_i, \sigma_x^2) \\
&\times \prod_{j=M+1}^{P} p(C_{ij}|C_{ij}^*) N(C_{ij}^*; \theta_0^j + \sum_{k=1}^{j-1} \alpha_k^j \theta_k^j C_{ik}^*, \sigma_j^2) \\
&\times P(\boldsymbol{\theta^y}) P(\boldsymbol{\theta^x}) P(\boldsymbol{\theta^j}) P(\sigma_y^2) P(\sigma_x^2) P(\boldsymbol{\sigma_j^2}) P(\boldsymbol{\alpha^y}, \boldsymbol{\alpha^x}) P(\boldsymbol{\alpha^j})
\end{aligned}
$$

Where $p(X_i|X_i^*) = \delta_{X_i^*}(X_i)$ for continuous X_i and $p(X_i|X_i^*) = X_i 1(X_i^* \geqslant 0) + (1 - X_i) 1(X_i^* < 0)$ for binary X_i. Analogous definitions hold for $p(Y_i|Y_i^*)$ and $p(C_{ij}|C_{ij}^*)$. For each regression coefficient in the model we assign independent, non-informative $N(0, K)$ priors, where K is set to be very large relative to the magnitude of the coefficients. For each variance parameter in the model we assign an $IG(a, b)$ prior. The prior distribution $P(\boldsymbol{\alpha^y}, \boldsymbol{\alpha^x})$ is implemented as described in the text. Under these priors the full conditionals take the following form:

$$P(\boldsymbol{\theta^x}|\bullet) \sim N\left(\left(\boldsymbol{W^x}(\boldsymbol{\alpha^x})^T\boldsymbol{W^x}(\boldsymbol{\alpha^x}) + \frac{\sigma_x^2 I}{k}\right)^{-1}\boldsymbol{W^x}(\boldsymbol{\alpha^x})^T\boldsymbol{X^*}, \left(\boldsymbol{W^x}(\boldsymbol{\alpha^x})^T\boldsymbol{W^x}(\boldsymbol{\alpha^x})/\sigma_x^2 + I/k\right)^{-1}\right)$$

$$P(\sigma_x^2|\bullet) \sim IG\left(N/2 + a, b + \frac{(\boldsymbol{X} - \boldsymbol{W^x}(\boldsymbol{\alpha^x})\boldsymbol{\theta^x})^T(\boldsymbol{X} - \boldsymbol{W^x}(\boldsymbol{\alpha^x})\boldsymbol{\theta^x})}{2}\right)$$

$$P(\boldsymbol{\theta^y}|\bullet) \sim N\left(\left(\boldsymbol{W^y}(\boldsymbol{\alpha^y})^T\boldsymbol{W^y}(\boldsymbol{\alpha^y}) + \frac{\sigma_y^2 I}{k}\right)^{-1}\boldsymbol{W^y}(\boldsymbol{\alpha^y})^T\boldsymbol{Y^*}, \left(\boldsymbol{W^y}(\boldsymbol{\alpha^y})^T\boldsymbol{W^y}(\boldsymbol{\alpha^y})/\sigma_y^2 + I/k\right)^{-1}\right)$$

$$P(\sigma_y^2|\bullet) \sim IG\left(N/2 + a, b + \frac{(\boldsymbol{Y} - \boldsymbol{W^y}(\boldsymbol{\alpha^y})\boldsymbol{\theta^y})^T(\boldsymbol{Y} - \boldsymbol{W^y}(\boldsymbol{\alpha^y})\boldsymbol{\theta^y})}{2}\right)$$

where $\boldsymbol{W^y}(\boldsymbol{\alpha^y})$ represents the design matrix for the outcome model defined by $\boldsymbol{\alpha^y}$, and $\boldsymbol{W^x}(\boldsymbol{\alpha^x})$ represents the design matrix for the exposure model defined by $\boldsymbol{\alpha^x}$. This means that the dimension of $\boldsymbol{\theta^x}$ and $\boldsymbol{\theta^y}$ are changing as we run through our gibbs sampler. In practice to implement this algorithm we set $\theta_j^y = 0$ when $\alpha_j^y = 0$ for all j, and then update the remaining values of $\boldsymbol{\theta^y}$ in the manner described above. We also note that if X or Y are binary then $\sigma_x^2 = 1$ or $\sigma_y^2 = 1$ by definition and no updating of those parameters is necessary. The full conditionals for the parameters of the imputation model for covariate j where j=M+1...P are as follows:

$$P(\boldsymbol{\theta^j}|\bullet) \sim N\left(\left(\boldsymbol{W^j}(\boldsymbol{\alpha^j})^T\boldsymbol{W^j}(\boldsymbol{\alpha^j}) + \frac{\sigma_j^2 I}{k}\right)^{-1}\boldsymbol{W^j}(\boldsymbol{\alpha^j})^T\boldsymbol{C_j^*}, \left(\boldsymbol{W^j}(\boldsymbol{\alpha^j})^T\boldsymbol{W^j}(\boldsymbol{\alpha^j})/\sigma_j^2 + I/k\right)^{-1}\right)$$

$$P(\sigma_j^2|\bullet) \sim IG\left(N/2 + a, b + \frac{(\boldsymbol{C_j} - \boldsymbol{W^j}(\boldsymbol{\alpha^j})\boldsymbol{\theta_j})^T(\boldsymbol{C_j} - \boldsymbol{W^j}(\boldsymbol{\alpha^j})\boldsymbol{\theta_j})}{2}\right)$$

Where if covariate j is binary then by definition $\sigma_j^2 = 1$ and no updating of the variance parameter is needed. $\boldsymbol{W^j}(\boldsymbol{\alpha^j})$ represents the design matrix defined by $\boldsymbol{\alpha^j}$. Now we need to update from the full conditional distribution of the missing covariates. If covariate j is missing and continuous then we will impute missing values for subject i from a Normal distribution:

$$N(V_{ij}\mu_{ij}, V_{ij})$$

Where

$$\mu_{ij} = \alpha_j^y \frac{Y_{i(-j)}\theta_j^y}{\sigma_y^2} + \alpha_j^x \frac{X_{i(-j)}\theta_j^x}{\sigma_x^2} + \frac{\tilde{\mu_{ij}}}{\sigma_j^2} + \sum_{k=j+1}^{P} \frac{\theta_j^k C_{ik(-j)}^*}{\sigma_k^2}$$
$$V_{ij} = \alpha_j^y \frac{\theta_j^{y\,2}}{\sigma_y^2} + \alpha_j^x \frac{\theta_j^{x\,2}}{\sigma_x^2} + \frac{1}{\sigma_j^2} + \sum_{k=j+1}^{P} \frac{\theta_j^{k\,2}}{\sigma_k^2}$$

And

$$Y_{i(-j)} = Y_i^* - \theta_0^y - \beta X_i - \sum_{l\neq j, l=1}^{P} C_{il}\theta_l^y$$
$$X_{i(-j)} = X_i^* - \theta_0^x - \sum_{l\neq j} C_{il}\theta_l^x$$
$$\tilde{\mu_{ij}} = \theta_0^j + \sum_{k=1}^{j-1} \theta_k^j C_{ik}^*$$
$$C_{ik(-j)}^* = C_{ik}^* - \theta_0^k - \sum_{l\neq j, l=1}^{k-1} C_{il}^*\theta_l^k$$

Notice that there are no interaction terms involving S_i because we are imputing the covariates for subjects in the main study for whom $S_i = 0$ and the interaction terms disappear. When C_{ij} is binary we will impute it's corresponding latent variable, C_{ij}^*. If C_{ij} is observed then we still update it's full conditional from

$$C_{ij}^* \sim C_{ij} TN_+(V_{ij}\mu_{ij}, V_{ij}) + (1 - C_{ij}) TN_-(V_{ij}\mu_{ij}, V_{ij})$$

Where TN_+ represents a truncated normal distribution that only assigns positive probability to the positive real line, and TN_- the same but only assigning mass to the negative real line.

$$\mu_{ij} = \frac{\tilde{\mu_{ij}}}{\sigma_j^2} + \sum_{k=j+1}^{P} \frac{\theta_j^k C^*_{ik(-j)}}{\sigma_k^2}$$
$$V_{ij} = \frac{1}{\sigma_j^2} + \sum_{k=j+1}^{P} \frac{{\theta_j^k}^2}{\sigma_k^2}$$

Where again

$$\tilde{\mu_{ij}} = \theta_0^j + \sum_{k=1}^{j-1} \theta_k^j C^*_{ik}$$
$$C^*_{ik(-j)} = C^*_{ik} - \theta_0^k - \sum_{l\neq j, l=1}^{k-1} C^*_{il}\theta_l^k$$

For binary variables that are missing, we again get a mixture of truncated normals, though we replace the binary indicator with the posterior probability that variable is 1 or 0 as follows:

$$C^*_{ij} \sim \pi_{ij} TN_+(V_{ij}\mu_{ij}, V_{ij}) + (1-\pi_{ij}) TN_-(V_{ij}\mu_{ij}, V_{ij})$$

With the probability defined as

$$\pi_{ij} = \frac{\Phi(\mu_{ij})\phi(Y_{i(-j)} - \theta_j^y)\phi(X_{i(-j)} - \theta_j^x)}{\Phi(\mu_{ij})\phi(Y_{i(-j)} - \theta_j^y)\phi(X_{i(-j)} - \theta_j^x) + (1-\Phi(\mu_{ij}))\phi(X_{i(-j)})\phi(Y_{i(-j)})}$$

The only parameters left to sample from are the vector of variable inclusion indicators $(\boldsymbol{\alpha^x}, \boldsymbol{\alpha^y}, \boldsymbol{\alpha^{M+1}}, ..., \boldsymbol{\alpha^P})$ and to do this we can follow the ideas of Wang *et al.* (2015) utilizing the MC^3 technique for searching a model space (Madigan *et al.* , 1994, 1995). We will illustrate how to sample from $P(\boldsymbol{\alpha^y}|\boldsymbol{\alpha^x}, \boldsymbol{D})$, however, the algorithm for sampling from $P(\boldsymbol{\alpha^x}|\boldsymbol{\alpha^y}, \boldsymbol{D})$ is analagous. We can define a neighborhood of $\boldsymbol{\alpha^y}$ to be the set of all outcome models with one covariate either added or removed from the model defined by $\boldsymbol{\alpha^y}$. If we are at iteration t of our

current Markov chain, and we are currently at the values $(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}, \boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}})$, then we randomly draw a model $\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(1)}}$ from the neighborhood of $\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}$ and we accept the new model with probability

$$\min\left\{1, \frac{P(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(1)}}|\boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}}, \boldsymbol{D})}{P(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}|\boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}}, \boldsymbol{D})} = \frac{P(\boldsymbol{Y}|\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(1)}}, \boldsymbol{X}, \boldsymbol{C})}{P(\boldsymbol{Y}|\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}, \boldsymbol{X}, \boldsymbol{C})} * \frac{P(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(1)}}|\boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}})}{P(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}|\boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}})}\right\}$$

Otherwise the chain stays at $\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}$. We are easily able to calculate $\frac{P(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(1)}}|\boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}})}{P(\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}|\boldsymbol{\alpha}^{\boldsymbol{x}}_{\boldsymbol{(0)}})}$ using our conditional prior specification from Section 2.2 of the main manuscript. To calculate the ratio of marginal likelihoods we can use the BIC approximation to the Bayes factor (Raftery, 1995) defined as

$$\frac{P(\boldsymbol{Y}|\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(1)}}, \boldsymbol{X}, \boldsymbol{C})}{P(\boldsymbol{Y}|\boldsymbol{\alpha}^{\boldsymbol{y}}_{\boldsymbol{(0)}}, \boldsymbol{X}, \boldsymbol{C})} \approx exp\left\{\frac{1}{2}(BIC_0 - BIC_1)\right\}$$

if ω is set to ∞, we can use the following algorithm to sample from the inclusion parameters:

1. at a given iteration of the MCMC we first update α^y_p conditional on α^x_p. If $\alpha^x_p = 1$ then we set $\alpha^y_p = 1$ with probability 1. If $\alpha^x_p = 0$ then we assign equal prior weight to both potential values of α^y_p and let the ratio of BICs determine the acceptance probability.

2. Then update α^x_p conditional on α^y_p. If $\alpha^y_p = 0$ then we set $\alpha^x_p = 0$ with probability 1. If $\alpha^y_p = 1$ then we assign equal weight a priori to both possible values of α^x_p and again let the ratio of BICs dictate the acceptance probability.

To sample from $P(\boldsymbol{\alpha^p}|\bullet)$ for $p = M + 1, ..., P$ we can again exploit the BIC approximation in the same manner as $\boldsymbol{\alpha^x}$ and $\boldsymbol{\alpha^y}$, only now we use a flat prior on the model space instead of the conditional prior we built for our exposure and outcome models.

B. Extended simulation results

In this section we show the results of additional simulations used to assess the ability of GBAC to adjust for confounding bias.

B.1 *Null treatment effect*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the treatment effect is 0 instead of 0.5. More specifically the models used to generate the data were as follows:

$$Y_i = 1500 + 0X_i + 0.2X_iS_i + 0.15C_{2i} + 0.15C_{6i} + 0.15C_{7i} + \epsilon_i \tag{B.1}$$

$$\Phi^{-1}(P(X_i = 1)) = -1 + 0.6C_{2i} + 0.6C_{6i} + 0.6C_{7i} \tag{B.2}$$

What we want to assess: To examine our approach when the true causal effect is zero in the main study

Results: Figure B.1 shows very similar results to those seen in the simulation of the original manuscript with the MSE of the proposed approach being the lowest, and the biggest gains are achieved for the smallest validation sample sizes.

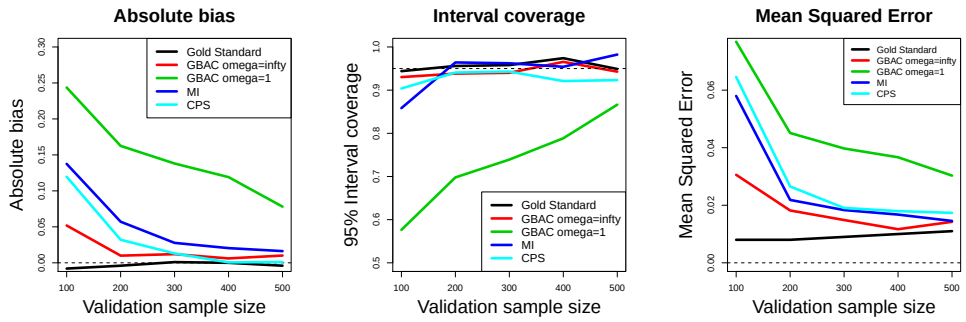

Fig. B.1. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations with a null treatment effect. $P = 50$, $M = 5$.

B.2 *Correlated data*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the covariates are correlated with each other. Specifically, the $P = 50$ covariates are drawn from a multivariate normal distribution with correlation 0.3 between each covariate.

What we want to assess: We want to test the proposed approach's ability to adjust for confounding when the covariates are correlated, since the simulation in the original manuscript included independent covariates only. Arguably, the covariates being independent is the most favorable setting for our approach in comparison with those that do not perform variable selection because it removes the unnecessary variables from the variable imputations. When every covariate is correlated with every other covariate, then this ability to remove noise variables is less important as each covariate should be included in the imputation models.

Results: Figure B.2 shows that we see similar results to that seen in the main manuscript as the proposed approach performs best with respect to MSE and achieves the desired interval coverage. We see that generally all of the approaches have slightly smaller MSE than in the manuscript, particularly at small validation sample sizes, because the added correlation gives us more data to impute the missing quantities with.

B.3 *Larger confounding*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the strength of confounding has been increased. To achieve this, we doubled the regression coefficients of the confounders from the original manuscript. This leads to data generating models that are as follows:

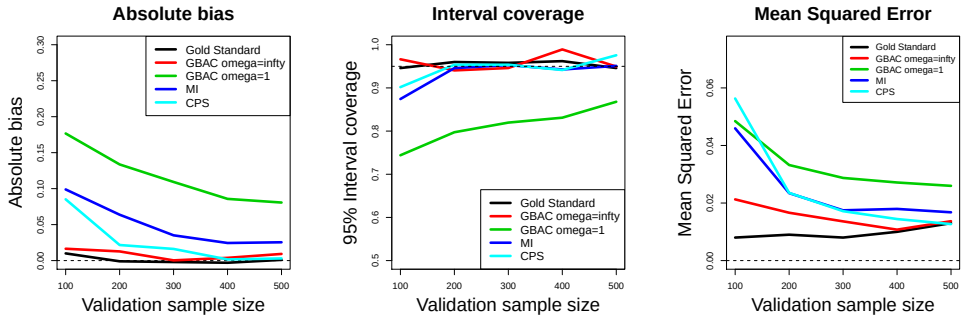

Fig. B.2. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations with correlated covariates. $P = 50$, $M = 5$.

$$Y_i = 1500 + 0X_i + 0.2X_iS_i + 0.3C_{2i} + 0.3C_{6i} + 0.3C_{7i} + \epsilon_i \tag{B.3}$$

$$\Phi^{-1}(P(X_i = 1)) = -1 + 1.2C_{2i} + 1.2C_{6i} + 1.2C_{7i} \tag{B.4}$$

What we want to assess: The ability of the proposed approach to adjust for confounding bias when the magnitude of confounding is larger than what was seen in the original manuscript.

Results: Figure B.3 shows that the proposed approach obtains the best MSE and the desired interval coverages. The disparity between the proposed approaches and the other approaches examined is larger than in the original manuscript. This is likely because the magnitude of the bias is larger and the proposed approach, as seen in the original manuscript, is better at adjusting for this bias.

B.4 *Smaller confounding*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the strength of confounding has been decreased. To achieve this, we halved the regression coefficients of the confounders from the original manuscript. This leads to data generating models that are as follows:

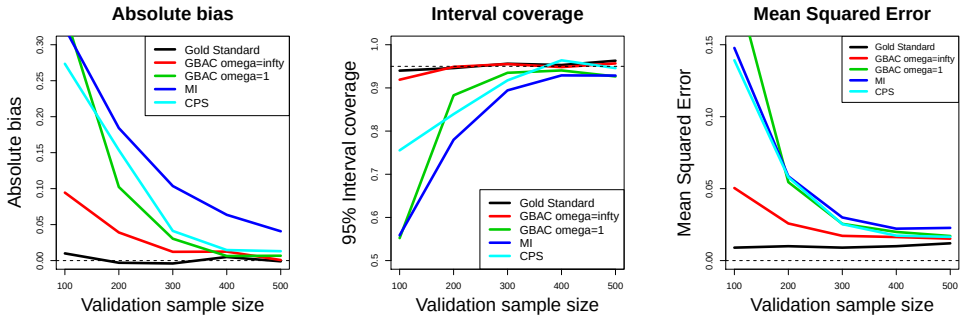

Fig. B.3. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations with larger confounding bias. $P = 50$, $M = 5$.

$$Y_i = 1500 + 0X_i + 0.2X_iS_i + 0.075C_{2i} + 0.075C_{6i} + 0.075C_{7i} + \epsilon_i \tag{B.5}$$

$$\Phi^{-1}(P(X_i = 1)) = -1 + 0.3C_{2i} + 0.3C_{6i} + 0.3C_{7i} \tag{B.6}$$

What we want to assess: The ability of the proposed approach to adjust for confounding bias when the magnitude of confounding is smaller than what was seen in the original manuscript.

Results: Figure B.4 shows that the proposed approach with ω set to infinity still does quite well in terms of MSE and interval coverage. The main difference in the results seen here and those seen in the manuscript is that GBAC(1) and the other approaches perform similarly well, particularly when the validation sample size is greater than 100. GBAC(1) actually performs slightly better than GBAC(∞) when the validation sample size is 100. This is likely because the confounding bias is so small that the differences are greatly mitigated in this scenario.

B.5 *Large and small confounding*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the strength of confounding has been changed for each confounder. The data generating models now are as follows:

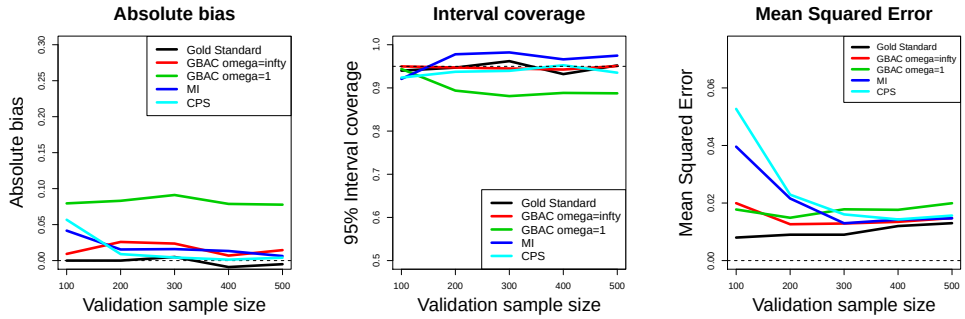

Fig. B.4. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations with smaller confounding bias. $P = 50$, $M = 5$.

$$Y_i = 1500 + 0X_i + 0.2X_iS_i + 0.15C_{2i} + 0.2C_{6i} - 0.4C_{7i} + \epsilon_i \tag{B.7}$$

$$\Phi^{-1}(P(X_i = 1)) = -1 + 0.6C_{2i} - 0.5C_{6i} - 0.8C_{7i} \tag{B.8}$$

What we want to assess: The ability of the proposed approach to adjust for confounding bias when the magnitude of confounding is different than what was seen in the original manuscript. This is simply a sanity check to make sure the situation in the manuscript is not one that just so happens to work well for our approach.

Results: Figure B.5 shows that the proposed approach with ω set to infinity still does quite well in terms of MSE and interval coverage. It outperforms the existing methods in bias and MSE for all validation sample sizes except $N_2 = 500$ when all of the approaches do quite well.

B.6 *Skewed missing data distribution for confounders*

Simulation setup: Instead of simulating independent, normal covariates, this simulation looks at the case where the confounders come from independent Gamma(3,3) distributions. It is important to note that only the true confounders come from this skewed distribution, while the other missing covariates are still truly normal.

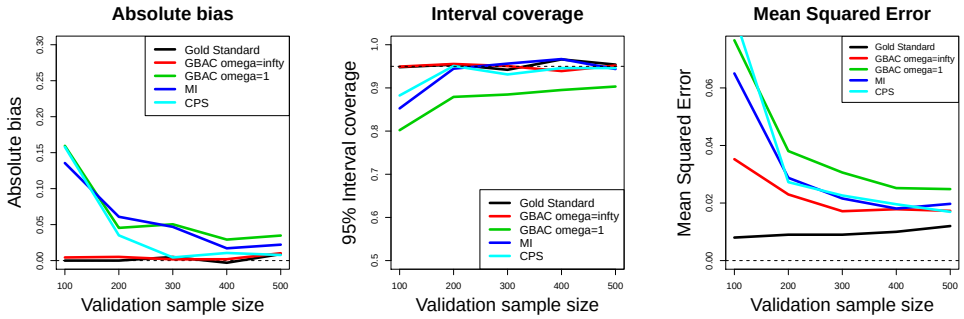

Fig. B.5. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations under different confounding bias than the original manuscript. $P = 50$, $M = 5$.

What we want to assess: The ability of the proposed approach to adjust for confounding bias when the distribution of the missing confounders is skewed and not normal, but the missing data model still assumes normality.

Results: Figure B.6 shows that the proposed approach is not able to eliminate bias for any validation sample size, which is expected due to the missing data model being incorrect. Importantly though, the bias is relatively small, and the overall MSE is not drastically worse than when the covariates are truly normally distributed. Due to the bias in the estimator, however, the interval coverage does not obtain the desired rate. Figure B.7 shows the inclusion probabilities for the $P = 50$ covariates in the study. We see that the inclusion of the important confounders (squares) is still very high using GBAC(∞) suggesting that misspecifying the distribution did not greatly affect the inclusion probabilities.

B.7 *Skewed missing data distribution for non-confounders*

Simulation setup: Instead of simulating independent, normal covariates, this simulation looks at the case where the missing covariates come from independent Gamma(3,3) distributions. It

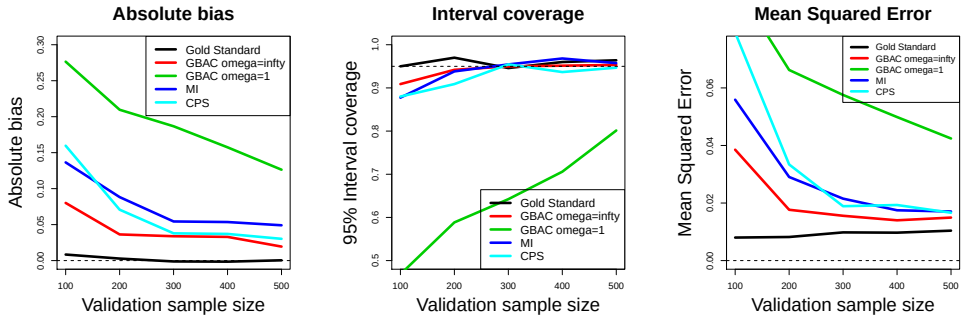

Fig. B.6. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations when the confounders come from a gamma(3,3) distribution. $P = 50$, $M = 5$.

is important to note that only the covariates which are not confounders come from this skewed distribution, while the true confounders are still truly normal.

What we want to assess: The ability of the proposed approach to adjust for confounding bias when the distribution of the missing covariates is skewed and not normal, but the missing data model still assumes normality. We also wish to assess how well our variable selection procedures do at removing noise variables from the model, when we've misspecified their missing data distributions.

Results: Figure B.8 shows that the relative performance of the proposed approach is unchanged when the true confounders are still normally distributed, but the noise variables are misspecified. This is intuitive when looking at Figure B.9, because we see that the noise variables are only included in the model a small percentage of the time, even though their distribution is misspecified.

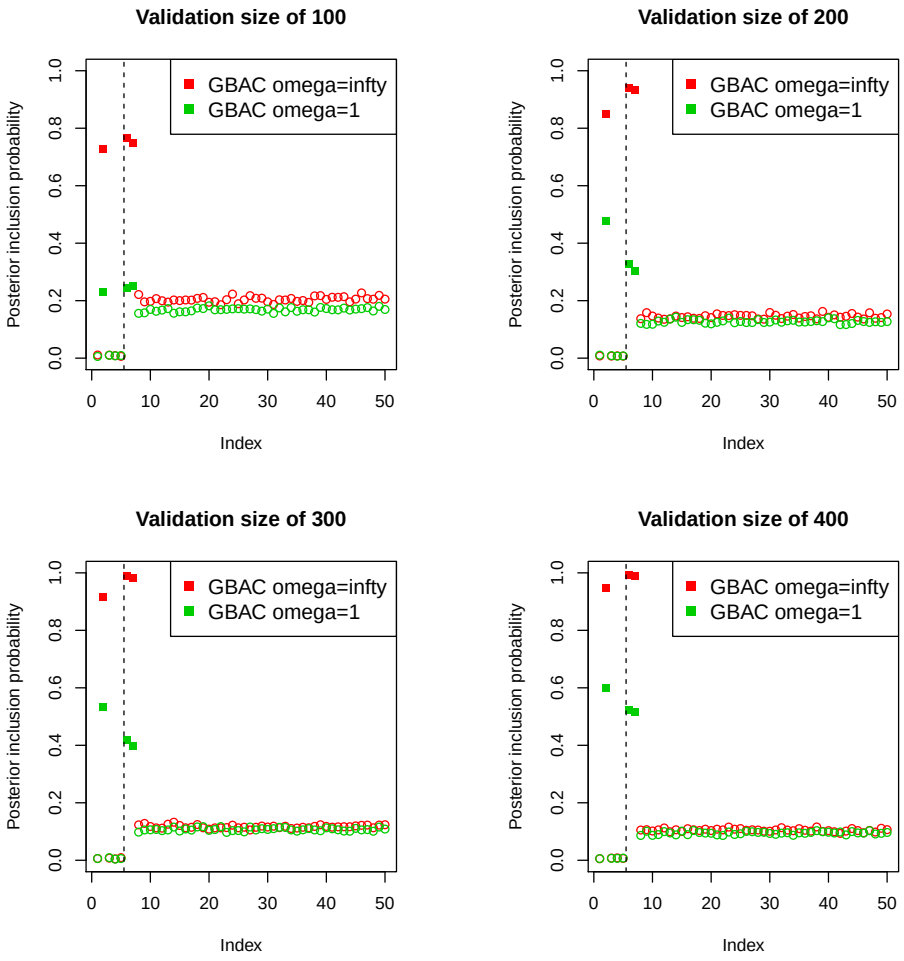

Fig. B.7. Inclusion probabilities for GBAC(1) and GBAC(∞) when the confounders come from a gamma(3,3) distribution.

B.8 *Misspecified exposure and outcome model*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the relationship between the confounders and both the exposure and outcome is non-linear and we assume them to be linear in our models. The data generating models now take the following form:

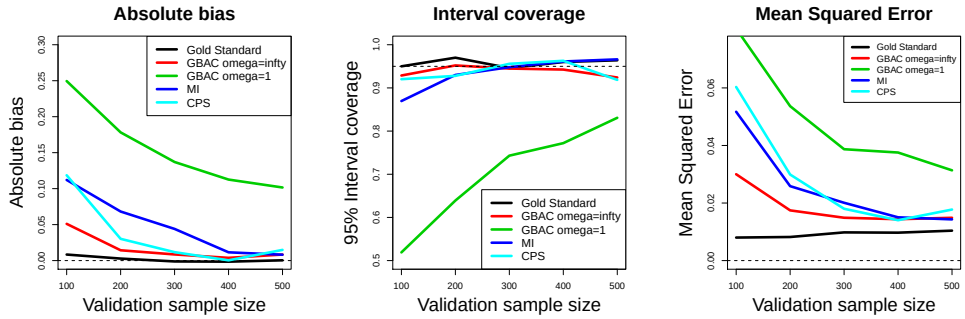

Fig. B.8. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations when the noise variables come from a gamma(3,3) distribution. $P = 50$, $M = 5$.

$$Y_i = 1500 + 0.5X_i + 0.2X_iS_i - 0.5C_2 + 0.4C_6^2 - 0.4e^{C_7} + \epsilon_i \tag{B.9}$$

$$\Phi^{-1}(P(X_i = 1)) = -1 + 0.2C_2^2 + 0.2e^{(0.4C_6)} + 0.3log(C_7^2) \tag{B.10}$$

What we want to assess: This simulation looks to evaluate the impact of grossly misspecifying the treatment and outcome models on the final causal effect estimates .

Results: Figure B.10 shows that all of the approaches being compared struggle when the true model is highly nonlinear. This is expected as all of the approaches assume linearity in one or both of these models. The proposed approach does the best in terms of MSE in this setting, however the bias and MSE of the proposed approach is far greater than when the true model is linear. Due to the large amount of bias in this case, interval coverage is compromised.

B.9 *Breaking assumption 3.8*

Simulation setup: This simulation is the same as the one included in the original manuscript, except now the data is no longer missing completely at random (MCAR) and the missingness mechanism leads to two data sets that no longer share the conditional distribution of the missing data conditional on the observed data. If we let $I_i = 1$ indicate that subject i was a member of

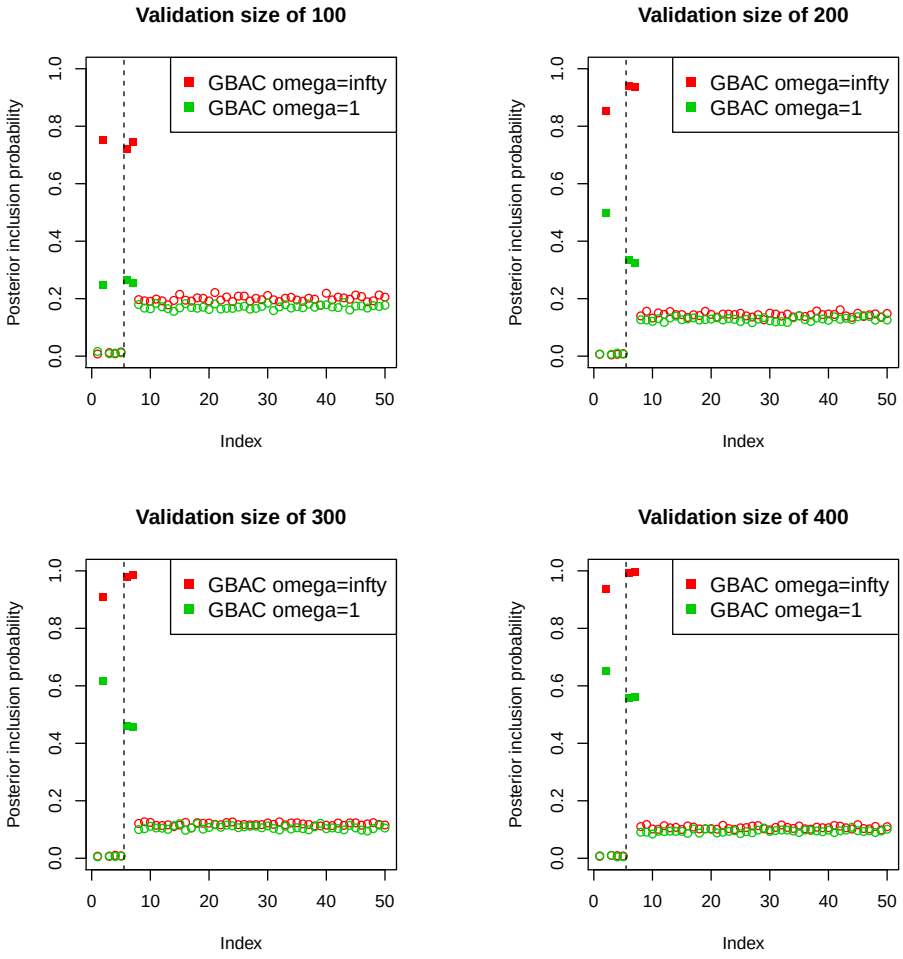

Fig. B.9. Inclusion probabilities for GBAC(1) and GBAC(∞) when the noise variables come from a gamma(3,3) distribution.

the main study and not the validation study then we assigned subjects to the main study via the following:

$$\log\left(\frac{p(I_i=1)}{1-p(I_i=1)}\right) \propto -1 + 0.25X_i + 0.5 \times 1(Y_i > \bar{y}). \tag{B.11}$$

This breaks our assumption that the distribution of the missing data conditional on everything else is the same between the main and validation data for a couple of reasons. For one, because

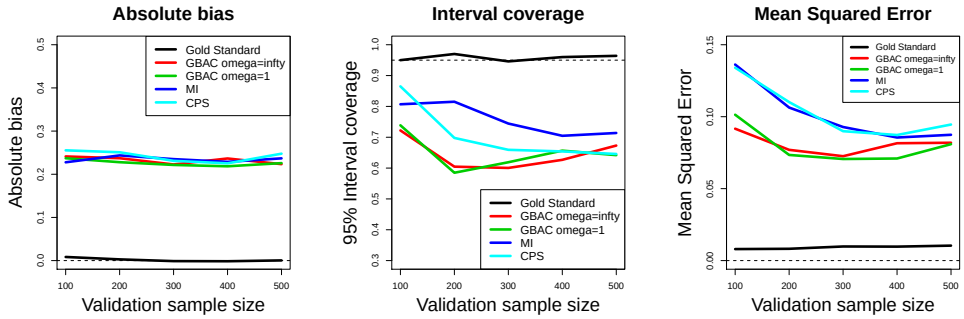

Fig. B.10. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations when the exposure and outcome models are nonlinear. $P = 50$, $M = 5$.

the treatment model is probit and this missing data mechanism will lead to different treatment rates in the two studies, the parameters of the treatment model will be different in the two studies due to non-collapsibility of probit models. Secondly, the distribution of the outcome conditional on everything else will be different in the two studies since those in the main study are more likely to have high outcome values relative to the overall mean. This is used to impute the missing data as well and will lead to different missing data distributions in the main and validation study.

What we want to assess: The sensitivity of our proposed approach to the assumption that the missing data distribution is shared between the two studies.

Results: Figure B.11 shows that our approach obtains some bias due to the lack of transportability between the two studies of the missing data distribution. This bias was not substantial, however, as the MSE of our approach isn't much larger than the simulation in the main manuscript where assumption 3.8 held. As in the main manuscript, our proposed approach greatly outperforms the other approaches in terms of bias and MSE.

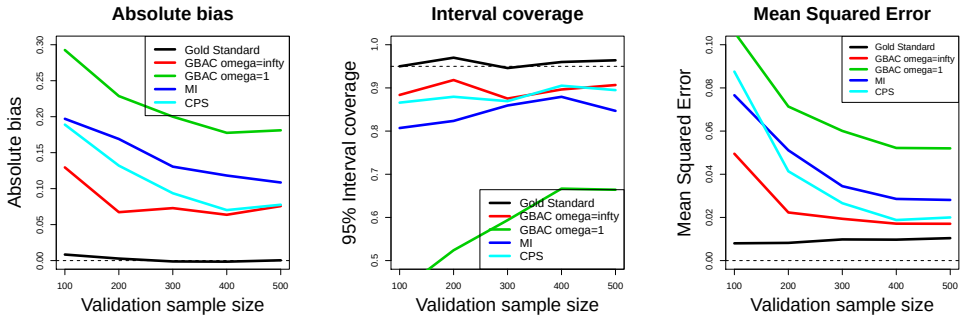

Fig. B.11. Bias, mean squared error, and interval coverage of the various estimators across 1000 simulations when the distribution of the missing data conditional on the observed data differs in the two studies. $P = 50$, $M = 5$.

References

Madigan, David, Raftery, Adrian E, York, Jeremy C, Bradshaw, Jeffrey M, & Almond, Russell G. 1994. Strategies for graphical model selection. *Pages 91–100 of: Selecting Models from Data*. Springer.

Madigan, David, York, Jeremy, & Allard, Denis. 1995. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, 215–232.

Raftery, Adrian E. 1995. Bayesian model selection in social research. *Sociological methodology*, **25**, 111–164.

Wang, Chi, Dominici, Francesca, Parmigiani, Giovanni, & Zigler, Corwin Matthew. 2015. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*.