Appendix

A Main Theorem

Our model is set up as follows:

1. For the ith subject, the true response Y_i depends on the covariates $\mathbf{X}_i$ following

$$Y_i \mid \mathbf{X} \sim \text{Bin}\left\{\frac{\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_0)}{1+\exp(\mathbf{X}_i^\top \boldsymbol{\beta}_0)} \mid \mathbf{X}\right\}.$$

2. The surrogate outcome S_i depends on Y_i following

$$\mathbb{P}\left(S_i = 1 \mid Y_i = 1\right) = p_1, \mathbb{P}\left(S_i = 0 \mid Y_i = 0\right) = p_0,$$

   where p_1 and p_0 are two constants.

3. To obtain the subset $\mathcal{V}$, the sampling procedure $\mathbb{I}(i \in \mathcal{V})$ depends on S_i following

$$\mathbb{P}(i \in \mathcal{V} \mid S_i = 1) = h_1, \mathbb{P}(i \in \mathcal{V} \mid S_i = 0) = h_0,$$

   where h_1 and h_0 are two preset constants.

Additionally, define $w_i = h_1 S_i + h_0(1 - S_i)$, the "biased" log-likelihood $l_i(\boldsymbol{\beta})$ and the "weighted" misspecified log-likelihood $g_i'(\boldsymbol{\gamma})$:

$$l_i(\boldsymbol{\beta}) = Y_i(c + \mathbf{X}_i\boldsymbol{\beta}) - \log\{1 + \exp(c + \mathbf{X}_i\boldsymbol{\beta})\} \text{ and } g_i'(\boldsymbol{\gamma}) = w_i\left[S_i\mathbf{X}_i\boldsymbol{\gamma} - \log\{1 + \exp(\mathbf{X}_i\boldsymbol{\gamma})\}\right]$$

Then

$$\nabla l_i(\boldsymbol{\beta}) = \left\{Y_i - \frac{\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})}{1+\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})}\right\}\mathbf{X}_i, \nabla^2 l_i(\boldsymbol{\beta}) = \left[\frac{\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})}{\left\{1+\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})\right\}^2}\right]\mathbf{X}_i\mathbf{X}_i^\top$$

$$\nabla g_i'(\boldsymbol{\gamma}) = w_i\left\{S_i - \frac{\exp(\mathbf{X}_i^\top\boldsymbol{\gamma})}{1+\exp(\mathbf{X}_i^\top\boldsymbol{\gamma})}\right\}\mathbf{X}_i, \nabla^2 g_i'(\boldsymbol{\gamma}) = w_i\frac{\exp(\mathbf{X}_i^\top\boldsymbol{\gamma})}{\{1+\exp(\mathbf{X}_i^\top\boldsymbol{\gamma})\}^2}\mathbf{X}_i\mathbf{X}_i^\top.$$

In addition to the model setting, we further require some regularity conditions to ensure the existence of our estimator.

Assumption 1.

1. $\mathbb{E}(\|\mathbf{X}_i\|_2^2) < \infty$ *for* $i = 1, \ldots, N$*;*
2. *The unique solutions to* $\mathbb{E}_{\boldsymbol{\beta}}\{\nabla l_i(\boldsymbol{\beta})\} = \mathbf{0}$ *and* $\mathbb{E}_{\boldsymbol{\gamma}}\{\nabla g'(S_i, \mathbf{X}_i; \boldsymbol{\gamma})\} = \mathbf{0}$ *exist.*

These assumptions are very mild, and hold in most of common models (linear models, generalized linear models, etc.).

Theorem 1. *For the MLE estimators* $\widehat{\boldsymbol{\beta}}_{\mathcal{V}}$*,* $\widehat{\boldsymbol{\gamma}}_{\mathcal{V}}$ *and* $\widehat{\boldsymbol{\gamma}}_{\mathcal{F}}$*, consider the combined estimator and the estimated variance*

$$\widehat{\boldsymbol{\beta}}_A = \widehat{\boldsymbol{\beta}}_{\mathcal{V}} - \widehat{\mathbf{H}}_Y^{-1}\widehat{\mathbf{G}}_{SY}\widehat{\mathbf{G}}_S^{-1}\widehat{\mathbf{H}}_S(\widehat{\boldsymbol{\gamma}}_{\mathcal{V}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{F}})$$

$$\widehat{var}(\widehat{\boldsymbol{\beta}}_A) = \widehat{\mathbf{H}}_Y^{-1} - (1-|\mathcal{V}||\mathcal{F}|^{-1})\widehat{\mathbf{H}}_Y^{-1}\widehat{\mathbf{G}}_{SY}\widehat{\mathbf{G}}_S^{-1}\widehat{\mathbf{G}}_{SY}^\top\widehat{\mathbf{H}}_Y^{-1},$$

where

$$\widehat{\mathbf{H}}_Y = |\mathcal{V}|^{-1}\sum_{i\in\mathcal{V}}\nabla^2 l_i(\widehat{\boldsymbol{\beta}}_{\mathcal{V}}), \qquad \widehat{\mathbf{H}}_S = |\mathcal{F}|^{-1}\sum_{i\in\mathcal{F}}\nabla^2 g_i'(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}})$$

$$\widehat{\mathbf{G}}_{SY} = |\mathcal{V}|^{-1}\sum_{i\in\mathcal{V}}\nabla l_i(\widehat{\boldsymbol{\beta}}_{\mathcal{V}})\nabla g_i'^\top(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}}), \qquad \widehat{\mathbf{G}}_S = |\mathcal{F}|^{-1}\sum_{i\in\mathcal{F}}\nabla g_i'(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}})\nabla g_i'^\top(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}}).$$

Then suppose $\widehat{\mathbf{H}}_Y^{-1}$ *and* $\widehat{\mathbf{G}}_S^{-1}$ *exist almost surely, under Assumption 1*

$$|\mathcal{V}|^{1/2}\{\widehat{var}(\widehat{\boldsymbol{\beta}}_A)\}^{-1/2}\left\{\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0 - \begin{pmatrix} c \\ \mathbf{0}\end{pmatrix}\right\} \rightsquigarrow N(\mathbf{0}, \mathbf{I}),$$

where c *is a constant.*

Proof.

Estimating Equation for S_i

Recall that the working estimation equation is $\nabla g(S_i, \mathbf{X}_i; \boldsymbol{\gamma})$. Then in the selected sub-samples, we are solving the new equation:

$$\mathbb{E}\{\nabla g(S_i, \mathbf{X}_i; \boldsymbol{\gamma})\mathbb{I}(i \in \mathcal{V})\} = \mathbf{0}$$

Suppose the solution $\boldsymbol{\gamma}_0'$ exists, denoted by $\widehat{\boldsymbol{\gamma}}_{\mathcal{F}}$ and $\widehat{\boldsymbol{\gamma}}_{\mathcal{V}}$ the empirical solutions to

$$|\mathcal{F}|^{-1}\sum_{i\in\mathcal{F}} \nabla g(S_i, \mathbf{X}_i; \boldsymbol{\gamma})\mathbb{P}(i \in \mathcal{V}) = 0 \text{ and } |\mathcal{V}|^{-1}\sum_{i\in\mathcal{V}} \nabla g(S_i, \mathbf{X}_i; \boldsymbol{\gamma}) = 0.$$

We have $\|\widehat{\boldsymbol{\gamma}}_{\mathcal{F}} - \boldsymbol{\gamma}_0'\| = o_{\mathrm{p}}(1)$ and $\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}} - \boldsymbol{\gamma}_0'\| = o_{\mathrm{p}}(1)$. Hence, for $i \in \mathcal{F}$, define the new estimating equation

$$\nabla g_i'(\boldsymbol{\gamma}) = \nabla g'(S_i, \mathbf{X}_i; \boldsymbol{\gamma}) = \nabla g(S_i, \mathbf{X}_i; \boldsymbol{\gamma})\mathbb{P}(i \in \mathcal{V}).$$

Specifically, in our case

$$\nabla g_i'(\boldsymbol{\gamma}_0) = w_i\left\{S_i - \frac{\exp(\mathbf{X}_i^\top\boldsymbol{\gamma}_0')}{1+\exp(\mathbf{X}_i^\top\boldsymbol{\gamma}_0')}\right\}\mathbf{X}_i, \nabla^2 g_i'(\boldsymbol{\gamma}_0) = w_i\frac{\exp(\mathbf{X}_i^\top\boldsymbol{\gamma}_0')}{\{1+\exp(\mathbf{X}_i^\top\boldsymbol{\gamma}_0')\}^2}\mathbf{X}_i\mathbf{X}_i^\top,$$

where $w_i = h_1 s_i + h_0(1 - S_i)$. Additionally, although not directly related, the detailed property regarding the misspecified models can be found in White (1982).

Biased Estimator for $\boldsymbol{\beta}_0$

On the other hand, given the biased validation set $\mathcal{V}$, for $\boldsymbol{\beta}_0$ estimation,

$$\begin{aligned}
&\mathbb{P}(Y_i = 1 \mid i \in \mathcal{V})\\
=&\frac{\mathbb{P}(Y_i = 1, i \in \mathcal{V})}{\mathbb{P}(i \in \mathcal{V})} = \frac{\mathbb{P}(Y_i = 1, S_i = 1, i \in \mathcal{V}) + \mathbb{P}(Y_i = 1, S_i = 0, i \in \mathcal{V})}{\mathbb{P}(i \in \mathcal{V})}\\
=&\frac{\mathbb{P}(Y_i = 1, S_i = 1)\mathbb{P}(i \in \mathcal{V} \mid \mathbf{S}_1) + \mathbb{P}(Y_i = 1, S_i = 0)\mathbb{P}(i \in \mathcal{V} \mid \mathbf{S}_0)}{\mathbb{P}(i \in \mathcal{V} \mid S_i = 1)\mathbb{P}(S_i = 1) + \mathbb{P}(i \in \mathcal{V} \mid S_i = 0)\mathbb{P}(S_i = 0)}\\
=&\frac{p_1 h_1 \mathbb{P}(Y_i = 1) + (1 - p_1)h_0\mathbb{P}(Y_i = 1)}{h_1\left[p_1\mathbb{P}(Y_i = 1) + (1 - p_0)\{1 - \mathbb{P}(Y_i = 1)\}\right] + h_0\left[p_0\{1 - \mathbb{P}(Y_i = 1)\} + (1 - p_1)\mathbb{P}(Y_i = 1)\right]}\\
=&\frac{\{p_1 h_1 + (1 - p_1)h_0\}\exp(\mathbf{X}_i^\top\boldsymbol{\beta}_0)}{h_1\{p_1\exp(\mathbf{X}_i^\top\boldsymbol{\beta}_0) + (1 - p_0)\} + h_0\{p_0 + (1 - p_1)\exp(\mathbf{X}_i^\top\boldsymbol{\beta}_0)\}}\\
=&\frac{\exp(c + \mathbf{X}_i^\top\boldsymbol{\beta}_0)}{1 + \exp(c + \mathbf{X}_i^\top\boldsymbol{\beta}_0)},
\end{aligned}$$

where $c = \log[\{h_1p_1 + h_0(1-p_1)\}\{h_1(1-p_0) + h_0p_0\}^{-1}]$. Let $l_i(\boldsymbol{\beta})$ be the log-likelihood of Y_i, then

$$\nabla l_i(\boldsymbol{\beta}) = \left\{Y_i - \frac{\exp(c + \mathbf{X}_i^\top\boldsymbol{\beta})}{1+\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})}\right\}\mathbf{X}_i, \quad \nabla^2 l_i(\boldsymbol{\beta}) = \left[\frac{\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})}{\left\{1+\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta})\right\}^2}\right]\mathbf{X}_i\mathbf{X}_i^\top$$

Therefore, we obtain the MLE estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{V}}$,

$$n^{1/2}\left\{\widehat{\boldsymbol{\beta}}_{\mathcal{V}} - \boldsymbol{\beta}_0 - \begin{pmatrix} c \\ \mathbf{0}\end{pmatrix}\right\} = n^{1/2}\{\mathbf{H}'_Y(\widehat{\boldsymbol{\beta}}_{\mathcal{V}})\}^{-1}\nabla\bar{l}_{\mathcal{V}}(\boldsymbol{\beta}_0) \rightsquigarrow N\{\mathbf{0}, \mathbf{G}_y^{-1}(\boldsymbol{\beta}_0)\},$$

where $\nabla\bar{l}_{\mathcal{V}}(\boldsymbol{\beta}) = |\mathcal{V}|^{-1}\sum_{i\in\mathcal{V}}\nabla l_i(\boldsymbol{\beta})$ and $\mathbf{H}'_Y(\widehat{\boldsymbol{\beta}}_{\mathcal{V}}) = \int_0^1 \nabla^2\bar{l}_{\mathcal{V}}\{\widehat{\boldsymbol{\beta}}_{\mathcal{V}} + t(\widehat{\boldsymbol{\beta}}_{\mathcal{V}} - \boldsymbol{\beta}_0)\}\mathrm{d}t$. Additionally, $\mathbf{G}_y(\boldsymbol{\beta}_0) = \mathbb{E}\{\nabla^2 l_i(\boldsymbol{\beta}_0)\}$ is the Fisher information matrix with respect to $\boldsymbol{\beta}_0$.

Combining all results, we can apply the formula in Chen and Chen (2000) and define

$$\widehat{\boldsymbol{\beta}}_A = \widehat{\boldsymbol{\beta}}_{\mathcal{V}} - \widehat{\mathbf{H}}_Y^{-1}\widehat{\mathbf{G}}_{SY}\widehat{\mathbf{G}}_S^{-1}\widehat{\mathbf{H}}_S(\widehat{\boldsymbol{\gamma}}_{\mathcal{V}} - \widehat{\boldsymbol{\gamma}}_{\mathcal{F}})$$
$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_A) = \widehat{\mathbf{H}}_Y^{-1} - (1 - |\mathcal{V}||\mathcal{F}|^{-1})\widehat{\mathbf{H}}_Y^{-1}\widehat{\mathbf{G}}_{SY}\widehat{\mathbf{G}}_S^{-1}\widehat{\mathbf{G}}_{SY}^\top\widehat{\mathbf{H}}_Y^{-1},$$

where

$$\widehat{\mathbf{H}}_Y = |\mathcal{V}|^{-1}\sum_{i\in\mathcal{V}}\nabla^2 l_i(\widehat{\boldsymbol{\beta}}_{\mathcal{V}}), \quad \widehat{\mathbf{H}}_S = |\mathcal{F}|^{-1}\sum_{i\in\mathcal{F}}\nabla^2 g'_i(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}})$$
$$\widehat{\mathbf{G}}_{SY} = |\mathcal{V}|^{-1}\sum_{i\in\mathcal{V}}\nabla l_i(\widehat{\boldsymbol{\beta}}_{\mathcal{V}})\nabla g_i'^\top(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}}), \quad \widehat{\mathbf{G}}_S = |\mathcal{F}|^{-1}\sum_{i\in\mathcal{F}}\nabla g'_i(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}})\nabla g_i'^\top(\widehat{\boldsymbol{\gamma}}_{\mathcal{F}}).$$

By the Slutsky's theorem, we have

$$|\mathcal{V}|^{1/2}\{\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_A)\}^{-1/2}\left\{\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0 - \begin{pmatrix} c \\ \mathbf{0}\end{pmatrix}\right\} \rightsquigarrow N(\mathbf{0}, \mathbf{I}).$$

□

B Group Assignment with Fixed Total Validation Size

Recall that our setting gives

$$\mathbb{P}(Y_i = 1 \mid i \in \mathcal{V}) = \frac{\exp(c + \mathbf{X}_i^\top\boldsymbol{\beta}_0)}{1+\exp(c+\mathbf{X}_i^\top\boldsymbol{\beta}_0)},$$

where $c = \log\left[\{h_1 p_1 + h_0(1-p_1)\}\{h_1(1-p_0)+h_0p_0\}^{-1}\right]$. Let ρ_0 be the marginal prevalence of Y. Under the null hypothesis,

$$\rho_0 = \frac{\exp(b_0)}{1+\exp(b_0)},$$

where b_0 is the intercept term. We hope $\mathbb{P}(Y_i = 1 \mid i \in \mathcal{V}) = 0.5$ in order to achieve a balanced comparison within $\mathcal{V}$. Therefore, $c + b_0 = 0$ and

$$\log\left\{\frac{h_1p_1 + h_0(1-p_1)}{h_1(1-p_0)+h_0p_0}\right\} + \log\left(\frac{\rho_0}{1-\rho_0}\right) = 0.$$

On the other hand, when the total validation size n is fixed, let $N_k = \sum_{i\in\mathcal{F}} \mathbb{I}(S_i = k)$ for $k = 1, 2$, $h_1 = n_1 N_1^{-1}$ and $h_0 = (n-n_1)N_0^{-1}$. Consequently,

$$\frac{n_1N_1^{-1}p_1 + (n-n_1)N_0^{-1}(1-p_1)}{n_1N_1^{-1}(1-p_0)+(n-n_1)N_0^{-1}p_0}\frac{\rho_0}{1-\rho_0} = 1$$
$$\frac{n_1\left\{N_1^{-1}p_1 - N_0^{-1}(1-p_1)\right\}\rho_0 + nN_0^{-1}(1-p_1)\rho_0}{n_1\left\{N_1^{-1}(1-p_0) - N_0^{-1}p_0\right\}(1-\rho_0) + nN_0^{-1}p_0(1-\rho_0)} = 1.$$

This leads to our conclusion:

$$n_1 = \frac{nN_1\{p_0(1-\rho_0) - (1-p_1)\rho_0\}}{\{N_0p_1 - N_1(1-p_1)\}\rho_0 - \{N_0(1-p_0) - N_1p_0\}(1-\rho_0)}. \tag{1}$$

However, in rare disease studies, often, ρ_0 is so low that the required n_1 goes far beyond n. In this case, we hope to make c as large as possible:

$$n_1 = \underset{n_1\in[0,n]}{\arg\max} \frac{n_1\{N_0p_1 - N_1(1-p_1)\} + nN_1(1-p_1)}{n_1\{N_0(1-p_0) - N_1p_0\} + nN_1p_0} \tag{2}$$

Especially, when both ρ_0 and $1-p_1 \to 0$, we can set $n_1 = n$ and $n_0 = 0$ to select as many cases as possible.

C Additional Simulation Results

This section provides additional simulation results of the model specified in Section 3.1 when $p_1 = 60\%$ and 80%. Box plots of MSE and empirical average coverage probabilities are presented.

Additional simulation results for the power comparisons with prevalence of 3%/5% are

Table 2: Empirical average coverage probabilities at 95% level for $p_1 = 60\%$. Different combinations of prevalence and specificity are demonstrated.

Prevalence (%)	p_0(%)	Oracle	Ori-Unif	Aug-Unif	Ori-Bias	OSCA
5	60	95	95	94	96	95
	80	95	95	95	95	95
	90	95	95	94	95	94
10	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	95	95	94
30	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	95	95	94
50	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	95	95	94

Table 3: Empirical average coverage probabilities at 95% level for $p_1 = 80\%$. Different combinations of prevalence and specificity are demonstrated.

Prevalence (%)	p_0(%)	Oracle	Ori-Unif	Aug-Unif	Ori-Bias	OSCA
5	60	95	95	94	95	95
	80	95	95	94	95	95
	90	95	95	94	95	94
10	60	95	95	95	95	95
	80	95	95	95	95	94
	90	95	95	94	95	94
30	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	95	95	95
50	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	95	95	95

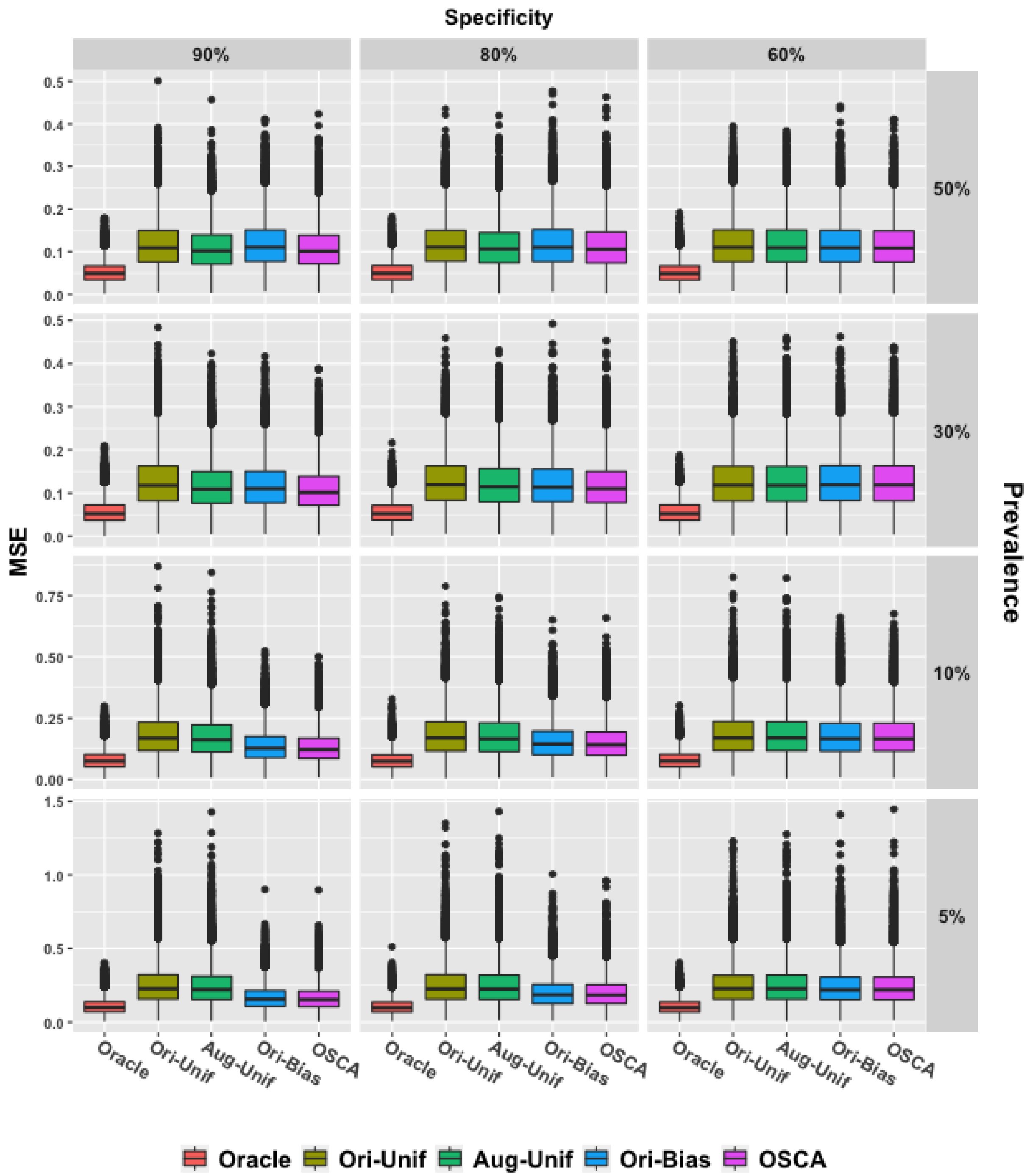

Figure 6: Box plots of the empirical MSE. Five methods are compared with fixed $p_1 = 60\%$. Each column gives results at different specificities (90%, 80% and 60%) and each row for different prevalence. The red, the gold, the green, the blue and the purple boxes respectively stand for the oracle method, the uniform-sampling method, the Aug-Unif method, the original biased-sampling method and the proposed method.

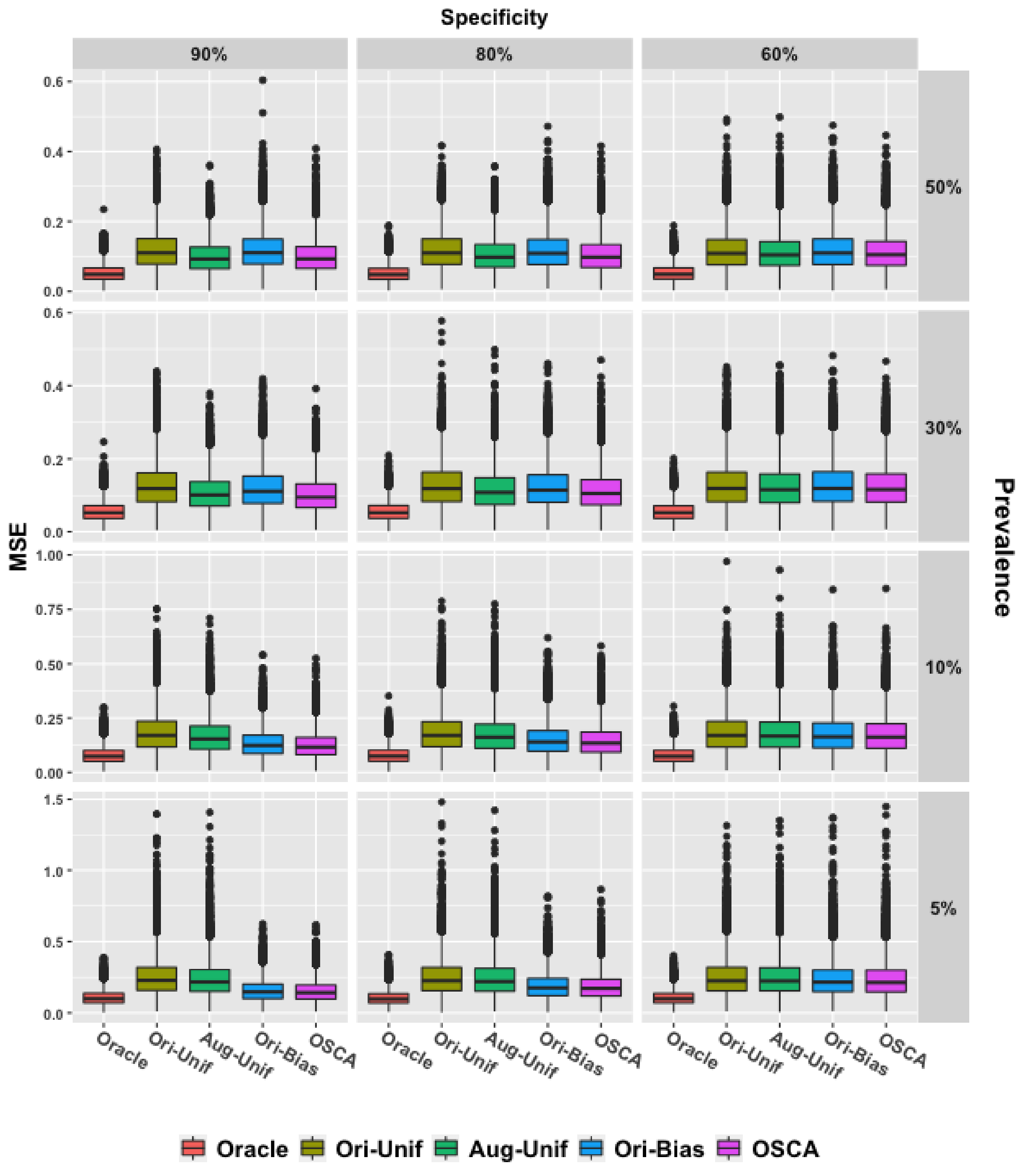

Figure 7: Box plots of the empirical MSE. Five methods are compared with fixed $p_1 = 80\%$. Each column gives results at different specificities (90%, 80% and 60%) and each row for different prevalence. The red, the gold, the green, the blue and the purple boxes respectively stand for the oracle method, the uniform-sampling method, the Aug-Unif method, the original biased-sampling method and the proposed method.

given in Figure 8.

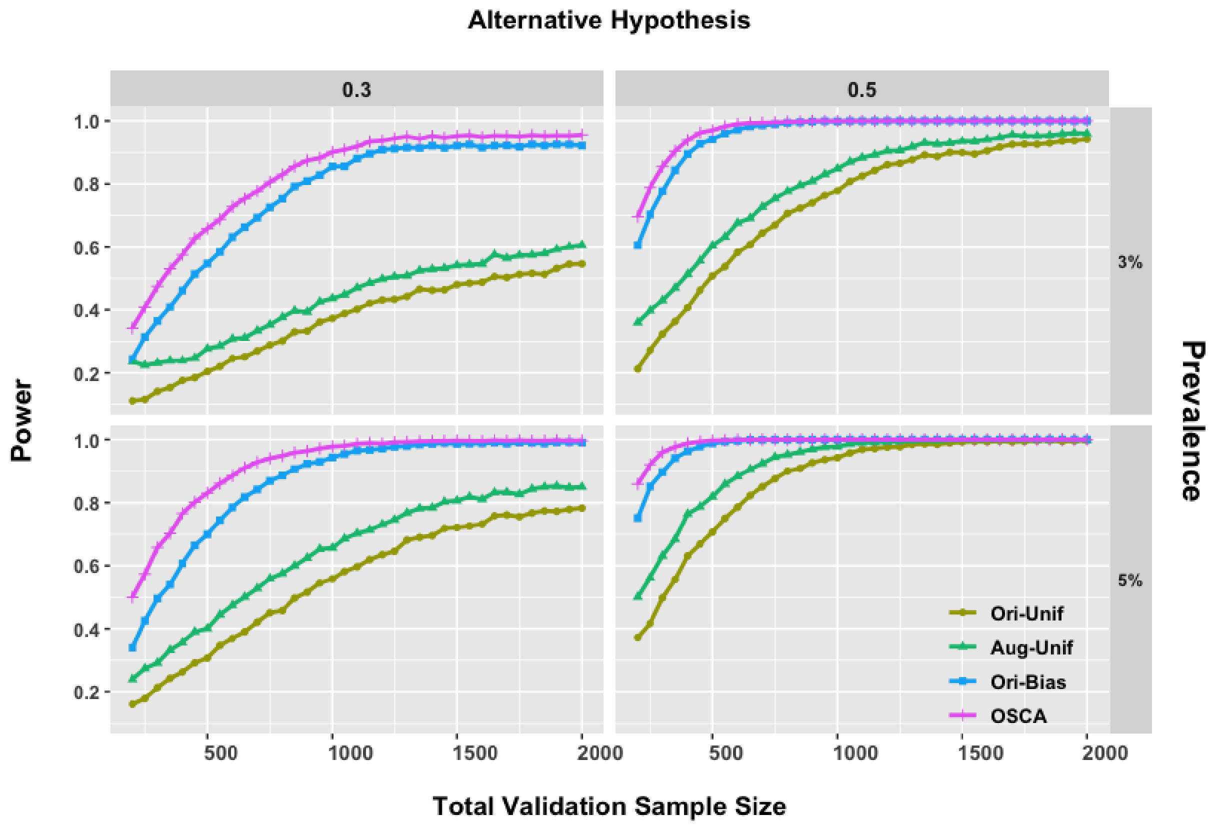

Figure 8: Power comparisons under different alternative hypotheses. Total validation sample size was varied from 200 to 2000. Combinations of prevalence at 3%/5% and alternative hypotheses $\beta_1 = 0.3/0.5$ were presented. In all panels, gold, green, blue, and purple lines stand for Ori-Unif, Aug-Unif, Ori-Bias and OSCA respectively.

References

Chen, Y. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(3):449–460.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

Table 4: Empirical average coverage probabilities at 95% level for $p_1 = 90\%$. Different combinations of prevalence and specificity are demonstrated.

Prevalence (%)	p_0(%)	Oracle	Ori-Unif	Aug-Unif	Ori-Bias	OSCA
5	60	95	95	95	96	95
	80	95	95	94	95	95
	90	95	95	94	96	94
10	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	94	95	94
30	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	94	95	94
50	60	95	95	95	95	95
	80	95	95	95	95	95
	90	95	95	95	95	95