Supplemental Material for

"Secondary outcome analysis for data from an outcome-dependent sampling design"

by Yinghao Pan, Jianwen Cai, Matthew Longnecker, and Haibo Zhou

1 Additional Simulation Studies

We conducted additional simulation studies, for which the set-up is similar to the Collaborative Perinatal Project data. In Table S1, we let the full cohort size be 40000, and then select an ODS sample with 850 SRS and 200 supplementals. The simulated covariates mimic the distribution of PCB levels and other confounding variables in the CPP data set.

Figure S1 is a histogram plot showing the difference between $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ in terms of estimating γ_2, where γ_2 describes the association between EDU and Y_2 (children's birth weight). Averaged across all the simulations, the IPW and AIPW estimates are very close. However, it is possible that in certain simulations, there exists observable difference between IPW and AIPW estimator. In fact, in 185 out of 1000 simulations, the absolute difference between IPW and AIPW estimator is larger than 10.

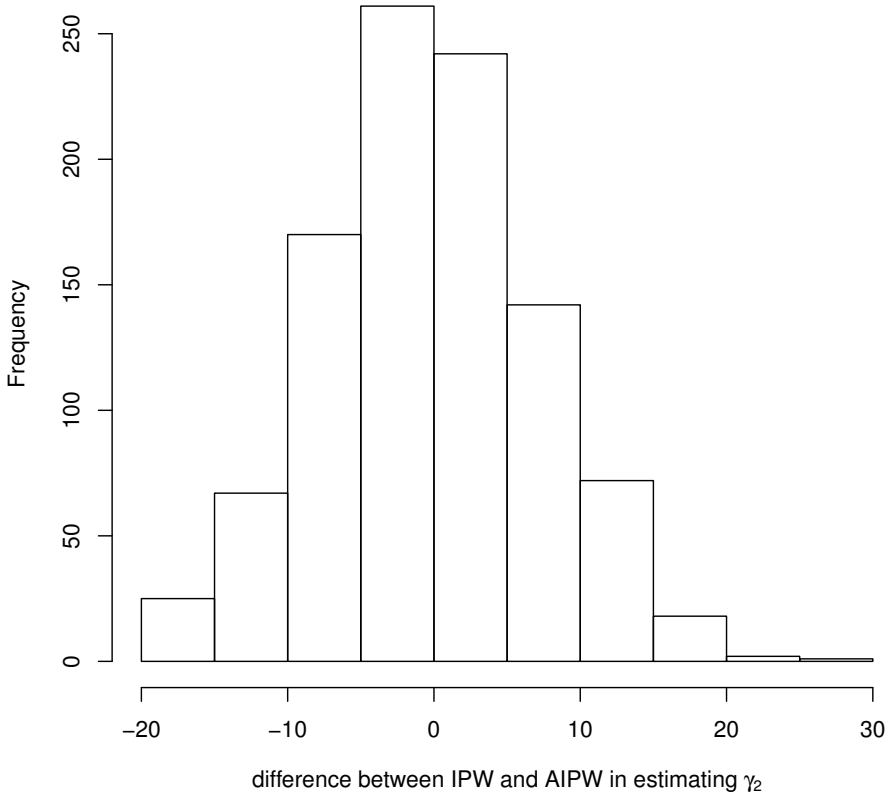

Figure S1: Histogram of the difference between $\hat{\xi}_{IPW}$ and $\hat{\xi}_{AIPW}$ with respect to estimating γ_2 in 1000 simulations, where γ_2 describes the association between EDU and Y_2 (children's birth weight).

Table S1: Simulation results based on 1000 simulations with $n_0 = 850$, $n_1 = n_3 = 100$, the validation sample size is $n_V = 1050$, the full cohort size is $N = 40000$.

Parameters	Methods	Mean	Var	$\widehat{Var}$	CI
γ_0	ξ_{SRS}	3099.07	7099.75	7091.51	0.955
	ξ_{IPW}	3097.82	5399.31	5643.01	0.946
	ξ_{AIPW}	3098.99	774.51	780.97	0.958
γ_1	ξ_{SRS}	−4.60	78.68	80.91	0.958
	ξ_{IPW}	−4.67	66.42	66.01	0.941
	ξ_{AIPW}	−4.69	66.91	66.00	0.944
γ_2	ξ_{SRS}	0.16	71.23	74.27	0.946
	ξ_{IPW}	0.27	57.46	59.84	0.950
	ξ_{AIPW}	0.04	1.59	1.76	0.963
γ_3	ξ_{SRS}	14.96	114.28	111.82	0.952
	ξ_{IPW}	14.79	94.87	92.07	0.947
	ξ_{AIPW}	14.93	2.69	2.68	0.951
γ_4	ξ_{SRS}	199.10	1521.78	1485.73	0.949
	ξ_{IPW}	200.52	1219.41	1247.56	0.955
	ξ_{AIPW}	200.12	36.48	35.75	0.948
γ_5	ξ_{SRS}	−123.19	1174.43	1185.49	0.954
	ξ_{IPW}	−122.11	944.33	977.70	0.961
	ξ_{AIPW}	−120.00	27.35	28.46	0.954

Results are based on the model $Y_1 = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + \beta_5 Z_4 + e$, $Y_2 = \gamma_0 + \gamma_1 X + \gamma_2 Z_1 + \gamma_3 Z_2 + \gamma_4 Z_3 + \gamma_5 Z_4 + \epsilon$, where (X, Z_1, Z_2, Z_3, Z_4) mimics the distribution of PCB, EDU, SES, RACE and GENDER from CPP data set. The error terms (e, ϵ) follow a bivariate normal distribution with $var(e) = \sigma_1^2$, $var(\epsilon) = \sigma_2^2$, $cov(e, \epsilon) = \rho\sigma_1\sigma_2$; the true parameter values are $\beta_0 = 72, \beta_1 = 0.25, \beta_2 = 1.2, \beta_3 = 1.5, \beta_4 = 9, \beta_5 = 0, \gamma_0 = 3100, \gamma_1 = -5, \gamma_2 = 0, \gamma_3 = 15, \gamma_4 = 200, \gamma_5 = -120, \sigma_1 = 12, \sigma_2 = 500, \rho = 0.5$. The cut off points for the ODS design are $\mu_{Y_1} \pm \sigma_{Y_1}$.