Supporting Document for Adjusted Logistic Propensity Weighting Methods for Population Inference using Nonprobability Volunteer-Based Epidemiologic Cohorts

By Lingxiao Wang, Richard Valliant, and Yan Li

A. Regularity Conditions

C1 The finite population size N, the cohort sample sizes n_c, and survey sample size n_s satisfy $\lim_{\substack{N\to\infty,\\ n_c\to\infty}} n_c/N = f_c \in (0,1)$, and $\lim_{\substack{N\to\infty,\\ n_p\to\infty}} n_p/N = f_p \in (0,1)$.

C2 There exist constants c_1 and c_2 such that $0 < c_1 \le N\pi_i^{(c)}/n_c \le c_2$, and $0 < c_1 \le N\pi_i^{(p)}/n_p \le c_2$ for all units $i \in F$.

C3 The finite population (FP) and the sample selection for s_s satisfy $N^{-1}\sum_{i\in s_p} d_i \boldsymbol{r}_i - N^{-1}\sum_{i\in FP} \boldsymbol{r}_i = O_p\left(n_p^{-1/2}\right)$, where $\boldsymbol{r}_i$ includes $\boldsymbol{x}_i$ and y_i where the order in probability is with respect to the probability sampling mechanism used to select s_p and $d_i = 1/\pi_i^{(p)}$.

C4 The FP and the propensity scores p_i's satisfy $N^{-1}\sum_{i\in FP} y_i^2 = O(1)$, $N^{-1}\sum_{i\in FP}\|\boldsymbol{x}_i\|^3 = O(1)$, $N^{-1}\sum_{i\in FP} p_i\boldsymbol{x}_i\boldsymbol{x}_i^T = O(1)$ being a positive definite matrix.

C5 The cohort participation and the survey sample selection satisfy $Cov\left(\delta_i^{(c)}, \delta_j^{(p)}\right) = 0$ for $i, j \in FP$.

Conditions **C1** – **C3** are regularly used in practice. Under **C1**, sample fractions of the nonprobability and probability sample are bounded. Condition **C2** indicates the (implicit) sample weights of nonprobability and probability sample units are bounded, i.e., $\pi_i^{(c)} = O(n_c/N)$ and $\pi_i^{(p)} = O\left(n_p/N\right)$, and the inclusion probabilities for the nonprobability and probability samples do not differ in terms of order of magnitude from simple random sampling. Condition **C3** guarantees consistency of the Horvitz-Thompson estimators obtained from the probability sample. Condition **C4** is the typical finite moment conditions to validate Taylor series expansions. Condition **C5** requires that selection of the nonprobability and the probability samples be independent, which simplifies the asymptotic variance calculation.

B. Proof of Theorem

We consider the following limiting process (Krewski & Rao, 1981; Chen, Li &Wu, 2019). Suppose there is a sequence of finite populations FP_k of size N_k, for $k = 1, 2, \cdots$. Cohort $s_{c,k}$ of size $n_{c,k}$ and survey sample $s_{p,k}$ of size $n_{p,k}$ are sampled from each FP_k. The sequences of the finite population, the cohort and the survey sample have their sizes satisfy $\lim_{k\to\infty} n_{t,k}/N_k \to f_t$ where $t = c$ or p and $0 < f_t \leq 1$ (regularity condition C1 in Appendix A). In the following the index k is suppressed for simplicity.

Let $\boldsymbol{\eta}^T = (\mu, \boldsymbol{\beta}^T)$. The ALP estimate of the finite population mean, $\hat{\mu}^{ALP}$, given in expression (2.3.6) in the main text, along with the estimates of propensity score model parameters, $\widehat{\boldsymbol{\beta}}$ (solution of $\tilde{S}^*(\boldsymbol{\beta}) = 0$ in expression (2.3.7) in the main text), can be combined as $\widehat{\boldsymbol{\eta}}^T = \left(\hat{\mu}^{ALP}, \widehat{\boldsymbol{\beta}}^T\right)$, which is the solution to the joint pseudo estimating equations

$$\Phi(\boldsymbol{\eta}) = \begin{pmatrix} U(\mu) = \frac{1}{N}\sum_{i\in FP} \delta_i^{(c)} \widetilde{w}_i (y_i - \mu) \\ \tilde{S}^*(\boldsymbol{\beta}) = \frac{1}{N+n_c}\sum_{i\in FP} \delta_i^{(c)}(1-p_i)\boldsymbol{x}_i - \frac{1}{N+n_c}\sum_{i\in FP} \delta_i^{(p)} d_i p_i \boldsymbol{x}_i \end{pmatrix} = \mathbf{0}, \quad \text{(B.1)}$$

where $\widetilde{w}_i = 1/\pi_i^{(c)} = (1-p_i)/p_i$. Under the joint randomization of the propensity model (i.e., self-selection of s_c) and the sampling design of s_s, we have $E\{\Phi(\boldsymbol{\eta}_0)\} = \mathbf{0}$, where $\boldsymbol{\eta}_0^T = (\mu_0, \boldsymbol{\beta}_0^T)$ with μ_0 and $\boldsymbol{\beta}_0$ being the true value of μ and $\boldsymbol{\beta}$ respectively. The consistency of $\widehat{\boldsymbol{\eta}}$ follows similar arguments to those in Chen, Li & Wu (2019) (which cited Section 3.2 of Tsiatis (2007)). Under the conditions **C1**-**C4**, we have $\Phi(\widehat{\boldsymbol{\eta}}) = \mathbf{0}$ By applying the first-order Taylor expansion, we have

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 \doteq [E\{\phi(\boldsymbol{\eta}_0)\}]^{-1}\Phi(\boldsymbol{\eta}_0), \quad \text{(B.2)}$$

where $E\{\phi(\boldsymbol{\eta})\} = E\left\{\frac{\partial\Phi(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\right\} = \begin{pmatrix} U_\mu & U_{\boldsymbol{\beta}} \\ \mathbf{0} & S_{\boldsymbol{\beta}} \end{pmatrix}$, and

$$U_\mu = E(\partial U/\partial\mu) = -\frac{1}{N}\sum_{i\in FP} \pi_i^{(c)} \widetilde{w}_i = -1,$$

$$U_{\boldsymbol{\beta}} = E(\partial U/\partial \boldsymbol{\beta}^T) = \frac{1}{N}\sum\nolimits_{i\in FP} \pi_i^{(c)}(y_i - \mu)\frac{\partial \widetilde{w}_i}{\partial \boldsymbol{\beta}^T} = -\frac{1}{N}\sum\nolimits_{i\in FP}(y_i - \mu)\boldsymbol{x}_i^T$$

$$S_{\boldsymbol{\beta}} = E\left(\partial \tilde{S}^*/\partial \boldsymbol{\beta}\right) = -\frac{1}{N+n_c}\sum\nolimits_{i\in FP} \pi_i^{(c)} \cdot p_i(1-p_i)\boldsymbol{x}_i\,\boldsymbol{x}_i^T - \frac{1}{N+n_c}\sum\nolimits_{i\in FP} p_i(1-p_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$$

$$= -\frac{1}{N+n_c}\sum\nolimits_{i\in FP} p_i\boldsymbol{x}_i\boldsymbol{x}_i^T \text{ (negative definite by condition } \mathbf{C4})$$

It follows that $\hat{\mu} = \mu_0 + O_p\left(n_c^{-1/2}\right)$, and

$$Var(\widehat{\boldsymbol{\eta}}) \doteq [E\{\phi(\boldsymbol{\eta}_0)\}]^{-1} Var\{\Phi(\boldsymbol{\eta}_0)\}[E\{\phi(\boldsymbol{\eta}_0)\}^T]^{-1}, \tag{B.3}$$

where $[E\{\phi(\boldsymbol{\eta})\}]^{-1} = \begin{pmatrix} -1 & \frac{N+n_c}{N}\boldsymbol{b}^T \\ \mathbf{0} & S_{\boldsymbol{\beta}}^{-1} \end{pmatrix}$, and $\boldsymbol{b}^T = \{\sum_{i\in FP}(y_i - \mu)\boldsymbol{x}_i^T\}\{\sum_{i\in FP} p_i\boldsymbol{x}_i\boldsymbol{x}_i^T\}^{-1}$. The middle part of (B.3), i.e., $Var\{\Phi(\boldsymbol{\eta}_0)\}$, can be calculated by partitioning $\Phi(\boldsymbol{\eta}) = \Phi_1 + \Phi_2$, where

$$\Phi_1 = \sum\nolimits_{i\in FP}\begin{Bmatrix} \frac{1}{N}\delta_i^{(c)}\widetilde{w}_i(y_i - \mu) \\ \frac{1}{N+n_c}\delta_i^{(c)}(1-p_i)\boldsymbol{x}_i \end{Bmatrix}, \Phi_2 = \frac{-1}{N+n_c}\sum\nolimits_{i\in FP}\begin{Bmatrix} 0 \\ \delta_i^{(p)}d_i p_i \boldsymbol{x}_i \end{Bmatrix}.$$

Notice that Φ_1 and Φ_2 are independent under condition **C5**, because Φ_1 only involves randomization of cohort participation while Φ_1 only involves survey sample selection. Hence, $Var\{\Phi(\boldsymbol{\eta}_0)\} = Var(\Phi_1) + Var(\Phi_2)$ where

$$Var(\Phi_1) = \sum\nolimits_{i\in FP} p_i(1-2p_i)\begin{Bmatrix} \frac{1}{N^2}(y_i - \mu)^2/p_i^2 & \frac{1}{N(N+n_c)}(y_i - \mu)\boldsymbol{x}_i^T/p_i \\ \frac{1}{N(N+n_c)}(y_i - \mu)\boldsymbol{x}_i/p_i & \frac{1}{(N+n_c)^2}\boldsymbol{x}_i\boldsymbol{x}_i^T \end{Bmatrix}$$

under the assumption of Poisson sampling of the nonprobability sample, and

$$Var(\Phi_2) = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{D} \end{pmatrix},$$

with $\boldsymbol{D}$ being the design-based variance-covariance matrix under the probability sampling design for sample s_s. For example, if survey sample is randomly selected by Poisson sampling, $\boldsymbol{D} = (N + n_c)^{-2} \sum_{i \in FP} (d_i - 1) p_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T$.

The finite population variance of $\hat{\mu}^{ALP}$ is the first diagonal element of $Var(\widehat{\boldsymbol{\eta}})$, and given by

$$Var(\hat{\mu}^{ALP}) = \begin{pmatrix} -1 & \boldsymbol{b}^T \end{pmatrix} \cdot \big(Var(\Phi_1) + Var(\Phi_2) \big) \cdot \begin{pmatrix} -1 \\ \boldsymbol{b} \end{pmatrix}$$

$$= N^{-2} \sum_{i \in FP} p_i (1 - 2p_i) \left\{ \frac{(y_i - \mu)}{p_i} - \boldsymbol{b}^T \boldsymbol{x}_i \right\}^2 + \boldsymbol{b}^T \boldsymbol{D} \boldsymbol{b}.$$

Note $p_i = P(i \in s_c^* | s_c^* \cup FP) \leq 1/2$.

C. Comparing Orders of Magnitude of $Var(\widehat{\mu}^{ALP})$ and $Var(\widehat{\mu}^{CLW})$

The pseudo-weighted nonprobability sample estimator of the population mean is written as

$$\hat{\mu} = \frac{1}{\sum_{i \in s_c} \widetilde{w}_i} \sum_{i \in s_c} \widetilde{w}_i y_i$$

where $\widetilde{w}_i$ is the pseudoweight w_i^{ALP} in the ALP estimator $\hat{\mu}^{ALP}$

$$w_i^{ALP} = \frac{1 - \hat{p}_i}{\hat{p}_i} = \exp^{-1}\big(\widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i\big)$$

or the pseudoweight w_i^{CLW} in the CLW estimator $\hat{\mu}^{CLW}$

$$w_i^{CLW} = \frac{1}{\hat{\pi}_i^{(c)}} = 1 + \exp^{-1}(\widehat{\boldsymbol{\gamma}}^T \boldsymbol{x}_i)$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ are solutions of pseudo estimation equations $\tilde{S}^*(\boldsymbol{\beta}) = 0$ and $\tilde{S}(\boldsymbol{\gamma}) = 0$ in formulae (2.3.7) and (2.2.7) in the main text, respectively.

According to the law of total variance, finite population variance of $\hat{\mu}$ can be written as

$$V(\hat{\mu}) = E_w[V_c(\hat{\mu} | \widetilde{\boldsymbol{w}})] + V_w[E_c(\hat{\mu} | \widetilde{\boldsymbol{w}})] \qquad \text{(C.1)}$$

where $\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, \dots, \widetilde{w}_N)$ is the vector of pseudo nonprobability sample weight for the finite population; E_w and V_w are with respect to the propensity model; V_c and E_c are with respect to the nonprobability sampling process, and we have

$$E_c(\hat{\mu}|\tilde{\boldsymbol{w}}) = \frac{\sum_{i\in FP}\pi_i^{(c)}\tilde{w}_i y_i}{\sum_{i\in FP}\pi_i^{(c)}\tilde{w}_i} + O(n_c^{-1}) \text{ and}$$

$$V_c(\hat{\mu}|\tilde{\boldsymbol{w}}) = \frac{\sum_{i\in FP}\pi_i^{(c)}\left(1-\pi_i^{(c)}\right)\tilde{w}_i^2\left(y_i - \frac{\sum_{i\in FP}\pi_i^{(c)}\tilde{w}_i\, y_i}{\sum_{i\in FP}\pi_i^{(c)}\tilde{w}_i}\right)^2}{\left(\sum_{i\in FP}\pi_i^{(c)}\tilde{w}_i\right)^2}$$

assuming Poisson sampling. The first term in (C.1), which is $E_w[V_c(\hat{\mu}|\tilde{\boldsymbol{w}})]$, has order $O(n_c^{-1})$ for both $\hat{\mu}^{ALP}$ and $\hat{\mu}^{CLW}$ under condition **C2**. The second term in (C.1) is approximately

$$V_w[E_c(\hat{\mu}|\tilde{\boldsymbol{w}})] \doteq \left(\frac{\partial E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\partial \tilde{\boldsymbol{w}}}\right) V(\tilde{\boldsymbol{w}}) \left(\frac{\partial E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\partial \tilde{\boldsymbol{w}}}\right)^T \tag{C.2}$$

The middle term in (C.2)is

$$V(\tilde{\boldsymbol{w}}) = \left(\frac{\partial \tilde{\boldsymbol{w}}}{\partial \widehat{\mathbf{B}}}\right) V\left(\widehat{\mathbf{B}}\right) \left(\frac{\partial \tilde{\boldsymbol{w}}}{\partial \widehat{\mathbf{B}}}\right)^T = \left\{\frac{\partial}{\partial \widehat{\mathbf{B}}} \exp^{-1}\left(\widehat{\mathbf{B}}^T \boldsymbol{x}\right)\right\}\left\{V\left(\widehat{\mathbf{B}}\right)\right\}\left\{\frac{\partial}{\partial \widehat{\mathbf{B}}} \exp^{-1}\left(\widehat{\mathbf{B}}^T \boldsymbol{x}\right)\right\}^T.$$

where $\widehat{\mathbf{B}} = \widehat{\boldsymbol{\beta}}$ or $\widehat{\boldsymbol{\gamma}}$ are solutions of pseudo estimating equations $\tilde{S}^*(\boldsymbol{\beta}) = 0$ and $\tilde{S}(\boldsymbol{\gamma}) = 0$ in the formulae (2.3.7) and (2.2.7). Therefore

$$V_w[E_c(\hat{\mu}|\tilde{\boldsymbol{w}})] \doteq \left(\frac{\partial E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\partial \tilde{\boldsymbol{w}}}\frac{\partial \tilde{\boldsymbol{w}}}{\partial \widehat{\mathbf{B}}}\right) V\left(\widehat{\mathbf{B}}\right) \left(\frac{\partial E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\partial \tilde{\boldsymbol{w}}}\frac{\partial \tilde{\boldsymbol{w}}}{\partial \widehat{\mathbf{B}}}\right)^T$$

where

$$\frac{\partial E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\partial \tilde{\boldsymbol{w}}} = \left\{\pi_1^{(c)}\frac{y_i - E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\sum_{i\in FP}\pi_1^{(c)}\tilde{w}_1}, \cdots, \pi_N^{(c)}\frac{y_i - E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\sum_{i\in FP}\pi_N^{(c)}\tilde{w}_N}\right\}^T,$$

and

$$\left(\frac{\partial E_c(\hat{\mu}|\tilde{\boldsymbol{w}})}{\partial \tilde{\boldsymbol{w}}}\frac{\partial \tilde{\boldsymbol{w}}}{\partial \widehat{\mathbf{B}}}\right) = -\frac{\sum_{i\in FP}\left\{\pi_i^{(c)}\exp^{-1}\left(\widehat{\mathbf{B}}^T\boldsymbol{x}_i\right)\left(y_i - E_c(\hat{\mu}|\tilde{\boldsymbol{w}})\right)\boldsymbol{x}_i\right\}}{\sum_{i\in FP}\pi_i^{(c)}\tilde{w}_i} = O(1)$$

for both ALP and CLW.

To solve the order of $V\left(\widehat{\mathbf{B}}\right)$, we first write

$$\widehat{\mathbf{B}} - \mathbf{B} = I^{-1}(\mathbf{B})S\left(\widehat{\mathbf{B}}\right) + o_p\left(S\left(\widehat{\mathbf{B}}\right)\right), \tag{C.3}$$

where $\boldsymbol{B} = \boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ are solutions to the census estimating equation $S(\mathbf{B}) = 0$, and $I(\mathbf{B}) = \frac{\partial S}{\partial \mathbf{B}}(\mathbf{B})$ is the Hessian matrix.

Specifically, for the ALP method the census estimating equation can be obtained by rewriting expression (3) in the main text and differentiating with respect to $\boldsymbol{\beta}$, leading to

$$S(\boldsymbol{\beta}) = \frac{1}{N+n_c}\sum\nolimits_{i\in s_c^* \cup FP}\{R_i - p_i(\boldsymbol{\beta})\}\boldsymbol{x}_i,$$

where R_i indicates the membership of s_c^* in $s_c^* \cup FP$ (=1 if $i \in s_c^*$; 0 if $i \in FP$), and $p_i(\boldsymbol{\beta}) = E(R_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) = \text{expit}(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ defined in Section 2.3 in the main text respectively.

The estimate $\widehat{\boldsymbol{\beta}}$ is solution to the pseudo estimating equation $\tilde{S}^*(\boldsymbol{\beta}) = 0$, where d_i is the basic design weights for $i \in s_p$ and $d_i = 1$ for $i \in s_c$. We have

$$\tilde{S}^*(\widehat{\boldsymbol{\beta}}) = \frac{1}{N+n_c}\sum\nolimits_{i\in s_c \cup^* s_p} d_i\{R_i - p_i(\widehat{\boldsymbol{\beta}})\}\boldsymbol{x}_i = S(\widehat{\boldsymbol{\beta}}) + O_p\left(\frac{1}{\sqrt{n_c+n_p}}\right) = 0,$$

under condition **C3**, where the union $\cup^*$ allows for duplicated units in s_c and s_p. Combined with (C.3), this leads to $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(n_c+n_p)^{-1/2}$ with

$$I(\boldsymbol{\beta}) = \frac{\partial S}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = -\frac{1}{N+n_c}\sum\nolimits_{i\in s_c^* \cup FP} p_i(\boldsymbol{\beta})\{1-p_i(\boldsymbol{\beta})\}\boldsymbol{x}_i = O(1)$$

under Condition **C4**. We have

$$V(\widehat{\boldsymbol{\beta}}) = O\left(\frac{1}{n_c+n_p}\right).$$

For the CLW method, the census estimating equation is

$$S(\boldsymbol{\gamma}) = \frac{1}{N}\sum\nolimits_{i\in FP}\left\{\delta_i - \pi_i^{(c)}(\boldsymbol{\gamma})\right\}\boldsymbol{x}_i$$

where δ_i is the indicator of the population unit i being included in s_c (=1 if $i \in s_c$; 0 otherwise), and $\pi_i(\boldsymbol{\gamma}) = E(\delta_i \mid \boldsymbol{x}_i; \boldsymbol{\gamma}) = \text{expit}(\boldsymbol{\gamma}^T\boldsymbol{x}_i)$.

The estimate $\widehat{\boldsymbol{\gamma}}$ is solution to the pseudo estimating equation $\tilde{S}(\boldsymbol{\gamma}) = 0$ shown below

$$\begin{aligned}\tilde{S}(\widehat{\boldsymbol{\gamma}}) &= \frac{1}{N}\left\{\sum\nolimits_{i\in s_c}\boldsymbol{x}_i - \sum\nolimits_{i\in s_p} d_i\pi_i^{(c)}(\widehat{\boldsymbol{\gamma}})\boldsymbol{x}_i\right\} \\ &= \frac{1}{N}\sum\nolimits_{i\in FP}\delta_i x_i + \frac{1}{N}\sum\nolimits_{i\in s_p} d_i\left\{\delta_i^{(c)} - \hat{\pi}_i^{(c)}\right\}\boldsymbol{x}_i - \frac{1}{N}\sum\nolimits_{i\in s_p} d_i\delta_i^{(c)}\boldsymbol{x}_i = 0.\end{aligned} \quad \text{(C.4)}$$

Under condition **C3**, we have the second and third term in (C.4)

$\frac{1}{N}\sum_{i\in s_p} d_i\left(\delta_i - \hat{\pi}_i^{(c)}\right)x_i = \frac{1}{N}\sum_{i\in FP}\left(\delta_i - \hat{\pi}_i^{(c)}\right)x_i + O_p\left(n_p^{-1/2}\right)$, and

$$\frac{1}{N}\sum_{i\in s_p} d_i\delta_i x_i = \frac{1}{N}\sum_{i\in FP}\delta_i x_i + O_p\left(n_p^{-1/2}\right).$$

Hence

$$\tilde{S}(\hat{\boldsymbol{\gamma}}) = S(\hat{\boldsymbol{\gamma}}) + O_p\left(n_p^{-1/2}\right) = 0,$$

which, combined with (C.3), leads to $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = O_p\left(n_p^{-1/2}\right)$ with

$$I(\boldsymbol{\gamma}) = -\frac{1}{N}\sum_{i\in FP}\pi_i^{(c)}(\boldsymbol{\gamma})\left\{1-\pi_i^{(c)}(\boldsymbol{\gamma})\right\}\boldsymbol{x}_i^T\boldsymbol{x}_i = O(1)$$

under condition **C6** in Chen, Li & Wu (2019).

We have

$$V(\hat{\boldsymbol{\gamma}}) = O\left(\frac{1}{n_p}\right)$$

As the result, the second term in (C.1) for the ALP and the CLW method has the order of $O\left(\frac{1}{n_p+n_c}\right)$ and $O\left(\frac{1}{n_p}\right)$, respectively. Combining the two terms in (C.1), we have

$$V(\hat{\mu}^{ALP}) = O\left(\frac{1}{n_p}\right) + O\left(\frac{1}{n_p+n_c}\right) = O\left(\frac{1}{n_c}\right)$$

and

$$V(\hat{\mu}^{CLW}) = O\left(\frac{1}{n_c}\right) + O\left(\frac{1}{n_p}\right) = O\left(\frac{1}{\min(n_c,n_p)}\right).$$

Therefore, in large samples we have $V(\hat{\mu}^{ALP}) \le V(\hat{\mu}^{CLW})$, and the estimator $\hat{\mu}^{ALP}$ is more efficient than $\hat{\mu}^{CLW}$ especially when $n_c \gg n_p$.

Notice that Comparison in analytical efficiency of the CLW and the ALP methods is made under their respective pseudo estimating equations (2.2.7) and (2.3.7) in Appendix C. Although the CLW pseudoweights are specified as $w_i^{CLW} = 1 + \exp^{-1}(\hat{\boldsymbol{\gamma}}^T\boldsymbol{x}_i)$, the justification also follows when $w_i^{CLW} = \exp^{-1}(\hat{\boldsymbol{\gamma}}^T\boldsymbol{x}_i)$. The ALP estimator tends to have smaller variance especially when the nonprobability sample is relatively larger than the probability sample, assuming nonprobability cohort and the survey sample are selected independently.

D. Supplementary table on estimated coefficients of propensity models

	RDW	CLW	ALP (FDW)	ALP.S
(Intercept)	-8.92	-8.92	-8.92	0.05

Age (in years)	-0.06	-0.06	-0.06	-0.06
Age^2	0.00	0.00	0.00	0.00
Sex (ref: male)				
Female	-0.10	-0.10	-0.10	-0.03
Education level	-0.16	-0.16	-0.16	-0.11
Race/Ethnicity (ref: NH-White)				
NH-Black	1.33	1.33	1.33	1.47
Hispanic	1.62	1.62	1.62	1.64
NH-Other	-0.35	-0.35	-0.35	-0.28
Poverty (ref: No)				
Yes	0.15	0.15	0.15	0.11
Unknown	-0.01	-0.01	-0.01	0.01
Health Status	0.24	0.24	0.24	0.24
Region (ref: Northeast)				
Midwest	0.25	0.25	0.25	0.15
South	0.41	0.41	0.41	0.35
West	0.29	0.29	0.29	0.14
Marital Status (ref: married or living as married)				
Single	-0.19	-0.19	-0.19	-0.12
Previously married	-0.01	-0.01	-0.01	-0.02
Smoking (ref: Non-smoker)				
Former smoker	0.12	0.12	0.12	0.10
Current smoker	0.16	0.16	0.16	0.14
Household Income	-0.01	-0.01	-0.01	-0.01
Chewing tobacco (ref: No)				
Yes	-0.35	-0.35	-0.35	-0.34
BMI (ref: normal)				
Under-weight	-0.02	-0.02	-0.02	-0.12
Over-weight	0.03	0.03	0.03	0.01
Obese	-0.06	-0.06	-0.06	-0.04

Reference

Chen, Y., Li, P., Wu, C. (2019) Doubly Robust inference with nonprobability survey samples. Journal of the American Statistical Association.; 1-11.

Krewski, D., Rao, J.N. (1981) Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. The Annals of Statistics, 1010-9.

Tsiatis, A. (2007). Semiparametric theory and missing data. Springer Science & Business Media.

Wang, L. (2020) Improving external validity of epidemiologic analyses by incorporating data from population-based surveys. Doctoral dissertation, University of Maryland, College Park.