Combining Probability and Non-Probability Samples

Michael Elliott[1,2]

[1]Department of Biostatistics, University of Michigan
[2]Survey Methodology Program, Institute for Social Research

Motivation for Utilizing Non-Probability Samples

- Non-probability samples are an increasing part of life for the survey analyst.
  - Non-response.
  - Sampling frame coverage.
  - Increasing cost.
  - Detailed outcomes of interest not present in probability samples.
  - Larger sample size than equivalent probability sample, especially in small domains.
- Offers possibility of improved inference if increase in precision is not overwhelmed by bias from the non-probability sample.

Framework for Nonprobability Sample Inference

Consider the joint density of a population vector of analysis variable $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$ and of 0-1 indicator variables $\delta_\mathbf{s} = (\delta_1, \delta_2, \ldots, \delta_N)$ for a sample s:

$$f(\mathbf{Y}, \delta_\mathbf{s} | \mathbf{X}; \Theta, \Phi) = f(\mathbf{Y} | \mathbf{X}; \Theta) f(\delta_\mathbf{s} | \mathbf{Y}, \mathbf{X}; \Phi)$$

where $\mathbf{X}$ is an $N \times p$ matrix of covariates that govern $\mathbf{Y}$ through unknown parameter Θ, and unknown parameter Φ governs $f(\delta_\mathbf{s}$ through both $\mathbf{Y}$ and $\mathbf{X}$ (Smith 1983; Rubin 1976; Little 1982).

- Probability sampling: $f(\delta_\mathbf{s} | \mathbf{Y}, \mathbf{X}; \Phi) = f(\delta_\mathbf{s} | \mathbf{X})$.
- Non-probability sampling: $\delta_\mathbf{s}$ can depend on $\mathbf{Y}$ and/or Φ in addition to $\mathbf{X}$.

Framework for Nonprobability Sample Inference

1. Quasi-randomization: model $f(\delta_{\mathbf{s}}|\mathbf{Y},\mathbf{X};\Phi)$.
   - Ideally, the probability of being in the sample is not NMAR and a model can be found for $f(\delta_{\mathbf{s}}|\mathbf{X};\Phi)$.
2. Superpopulation: model $f(\mathbf{Y}|\mathbf{X};\Theta)$.
   - Calibration a broad special case where model-based estimates are adjusted to known or estimated quantities outside of the non-probability sample.
3. Doubly robust models combine 1. and 2.
   - Extends the idea of augmented inverse propensity weighting: combines predicted means from models for probability sample with QR-weighted residuals from non-probability sample.

Key Assumptions

- Positivity: $P(\delta_i^B = 1 \mid \mathbf{x}_i) > 0$ for all $\mathbf{x}_i$.
- Quasi-Randomization
  - Ignorability: $Y_i \perp \delta_i^B \mid \mathbf{x}_i$.
  - Independence: $\delta_i^R \perp \delta_i^B \mid \mathbf{x}_i$.
- Superpopulation
  - Know $f(Y_i \mid \mathbf{x_i})$.

Quasi-Randomization

Population

	X	Y	δ^B	δ^R	π^B	π^R
S_B			1	0	?	?
			1	0	?	?
			⋮	⋮	⋮	⋮
			1	0	?	?
S_R		?	0	1	?	
		?	0	1	?	
		⋮	⋮	⋮	⋮	
		?	0	1	?	
$U-S$	?	?	0	0	?	?
	?	?	0	0	?	?
	⋮	⋮	⋮	⋮	⋮	⋮
	?	?	0	0	?	?

QR Complete data

	X	Y	δ^B	δ^R	π^B	π^R
S_B			1	0		
			1	0		
			⋮	⋮		
			1	0		

$$\overline{y}_{QR} = \frac{\sum_{i \in S_B} y_i (\hat{\pi}_i^B)^{-1}}{\sum_{i \in S_B} (\hat{\pi}_i^B)^{-1}}$$

Quasi-randomization: Generating Pseudo-Weights

- Elliott and Davis (2005) developed method to account for non-response bias and frame coverage.
  - Extend to estimate over- and under-representation of sample elements in the non-probability sample based on covariates available in both samples.
- By repeated application of Bayes' Rule we can estimate the probability that a nonprobability case would have been sampled by

$$\pi_i^B = P(\delta_i^B = 1 \mid \mathbf{x}_i = \mathbf{x}_o) = \frac{P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^B = 1)P(\delta_i^B = 1)}{P(\mathbf{x}_i = \mathbf{x}_o)}$$

$$= \frac{P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^B = 1)P(\delta_i^B = 1)P(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}{P(\delta_i^R = 1)P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^R = 1)}$$

$$\propto \frac{P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^B = 1)P(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}{P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^R = 1)}.$$

Generating Pseudo-Weights

- Estimating $P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^B = 1)$ and $P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^R = 1)$ can be difficult for a general joint distribution of covariates $\mathbf{x}$.
- Extensions of discriminant analysis provide a way around this problem. Let Z_i be an indicator for whether the subject is in the nonprobability sample.
- If sampling fractions are small

$$P(\delta_i^R = 1, \delta_i^B = 0) \approx P(\delta_i^R = 1) \text{ and } P(\delta_i^R = 0, \delta_i^B = 1) \approx P(\delta_i^B = 1) \text{ so}$$
$$P(\mathbf{x}_i | Z_i = 0) = P(\mathbf{x}_i | \delta_i^R = 1, \delta_i^B = 0) \approx P(\mathbf{x}_i | \delta_i^R = 1) \text{ and}$$
$$P(\mathbf{x}_i | Z_i = 1) = P(\mathbf{x}_i | \delta_i^R = 0, \delta_i^B = 1) \approx P(\mathbf{x}_i | \delta_i^B = 1).$$

Then

$$\frac{P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^B = 1)}{P(\mathbf{x}_i = \mathbf{x}_o \mid \delta_i^R = 1)} \approx \frac{P(\mathbf{x}_i = \mathbf{x}_o \mid Z_i = 1)}{P(\mathbf{x}_i = \mathbf{x}_o \mid Z_i = 0)}$$
$$= \frac{P(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o) P(\mathbf{x}_i = \mathbf{x}_o) / P(Z_i = 1)}{P(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_o) P(\mathbf{x}_i = \mathbf{x}_o) / P(Z_i = 0)}$$
$$\propto \frac{P(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}{P(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_o)}.$$

Generating Pseudo-Weights

- Resulting pseudo-weight is given by

$$w_i = 1/\hat{\pi}_i^B = 1/\hat{P}(\delta_i^B = 1 \mid \mathbf{x}_i = \mathbf{x}_o) \propto$$

$$1/\hat{P}(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)\frac{\hat{P}(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_o)}{\hat{P}(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}.$$

- If the probability sample weight as a function of $\mathbf{x}_o$ is known, $1/\hat{P}(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ can be replaced with known $1/\pi_i^R$.
  - Otherwise $\hat{P}(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ can be estimated using, e.g., beta regression (Ferrari and Cribari 2004).
- Obtain $\hat{P}(Z_i = z \mid \mathbf{x}_i = \mathbf{x}_o)$ via logistic regression.
  - LASSO (Tibshirani 1996).
  - Super learner algorithms (Van der Laan et al. 2007).
  - Bayesian additive regression trees (Chipman et al. 2010).

Generating Pseudo-Weights

- Resulting pseudo-weight is given by

$$w_i = 1/\hat{\pi}_i^B = 1/\hat{P}(\delta_i^B = 1 \mid \mathbf{x}_i = \mathbf{x}_o) \propto$$

$$1/\hat{P}(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)\frac{\hat{P}(Z_i = 0 \mid \mathbf{x}_i = \mathbf{x}_o)}{\hat{P}(Z_i = 1 \mid \mathbf{x}_i = \mathbf{x}_o)}.$$

- If the probability sample weight as a function of $\mathbf{x}_o$ is known, $1/\hat{P}(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ can be replaced with known $1/\pi_i^R$.
  - Otherwise $\hat{P}(\delta_i^R = 1 \mid \mathbf{x}_i = \mathbf{x}_o)$ can be estimated using, e.g., beta regression (Ferrari and Cribari 2004).
- Obtain $\hat{P}(Z_i = z \mid \mathbf{x}_i = \mathbf{x}_o)$ via logistic regression.
  - LASSO (Tibshirani 1996).
  - Super learner algorithms (Van der Laan et al. 2007).
  - Bayesian additive regression trees (Chipman et al. 2010).

Inference Under Quasi-Randomization

- $E(\hat{\bar{y}}_{QR}) = E(E(\hat{\bar{y}}_{QR} \mid \hat{\pi}^B))$
- $V(\hat{\bar{y}}_{QR}) = E(V(\hat{\bar{y}}_{QR} \mid \hat{\pi}^B)) + E(V(\hat{\bar{y}}_{QR} \mid \hat{\pi}^B))$
- Compute using Rubin's MI combining rule (Rafei et al. 2020): for each draw of $(\pi^B)^{(b)}$ from BART compute $\hat{\bar{y}}_{QR}^{(b)} = \frac{\sum_{i \in s_B} y_i ((\pi_i^B)^{(b)})^{-1}}{\sum_{i \in s_B} ((\pi_i^B)^{(b)})^{-1}}$ and
  $\hat{V}(\hat{\bar{y}}_{QR}^{(b)}) = \frac{N+1}{N} \frac{\sum_{i \in s_B} (y_i - \bar{y}_{QR}^{(b)})^2 ((\pi_i^B)^{(b)})^{-1}}{(\sum_{i \in s_B} ((\pi_i^B)^{(b)})^{-1})^2}$
- Then point and variance estimates are given by

$$\hat{\bar{y}}_{QR} = B^{-1} \sum_b \bar{y}_{QR}^{(b)}$$

$$\hat{V}(\bar{y}_{QR}) = B^{-1} \sum_b \hat{V}(\bar{y}_{QR}^{(b)}) + \frac{B+1}{B} (B-1)^{-1} \sum_b (\bar{y}_{QR}^{(b)} - \hat{\bar{y}}_{QR})^2$$

Simulation Study

- Generate population of 100,000 starting with design variables D and common covariates X:

$$\begin{pmatrix} D_1 \\ D_2 \\ X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & -\rho/2 & \rho & -\rho/2 \\ -\rho/2 & 1 & -\rho/2 & \rho \\ \rho & -\rho/2 & 1 & -\rho/2 \\ -\rho/2 & \rho & -\rho/2 & 1 \end{pmatrix} \right)$$

- Generate outcome given covariates

$$Y_i \mid x_i \sim N(-2 + x_{1i} - 2x_{2i} + 3x_{1i}x_{2i}, 1)$$

- Selection probabilities:

$$P(\delta_i^R = 1 \mid d_i) = \frac{e^{-1-0.5d_{1i}^2-d_{2i}}}{4(1+e^{-1-0.5d_{1i}^2-d_{2i}})}$$

$$P(\delta_i^B = 1 \mid d_i) = \frac{e^{-3-x_{1i}+x_{2i}-0.5x_{1i}x_{2i}}}{2(1+e^{-3-x_{1i}+x_{2i}-0.5x_{1i}x_{2i}})}$$

Simulation Study

- Given X Sample design is ignorable in S_B, not in S_R.
- Assume δ_i^R known only in the probability sample, δ_i^B unknown.
- Probability sample $n_R = 200$, non-probability sample $n_B = 1,000$.
- Assume $\rho = 0.8$.
- Consider alternatives:
  - Valliant and Dever (2018, p. 574) compute a weighted logistic regression to estimate $\pi_i^B = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$ by solving

$$U(\beta) = \sum_{i=1}^{n_B} x_i (1 - \pi_i^B(x_i, \beta)) - \sum_{i=1}^{n_R} x_i \pi_i^B(x_i, \beta)/\pi_i^R$$

  - Chen et al. (2020) replace population term in long-likelihood function with a Horvitz-Thompson type estimator to solve

$$U(\beta) = \sum_{i=1}^{n_B} x_i - \sum_{i=1}^{n_R} x_i \pi_i^B(x_i, \beta)/\pi_i^R$$

Simulation Study

Method	Rel Bias	RMSE	95% Cov	SE Ratio
Unweighted	125.6	125.8	0	0.98
Weighted	0.2	7.8	93.9	0.96
Valliant-Dever	-17.0	43.8	87.0	0.98
Chen et al.	-26.0	57.9	88.1	1.05
Quasi-rand:				
GLM	-15.1	22.0	82.7	1.01
BART	-3.9	14.6	96.9	1.04

Estimation Under Doubly Robust Estimation

- Can generate estimates of the selection probabilities π_i^B as in QR approach.
- Now need estimates of $\hat{y}_i$ based on available x_i
  - Can generate parametrically, or using BART.

Inference Under Doubly Robust Estimation

- Selection probabilities π^R known for nonprobability sample (follows derivation in Chen, Li, and Wu 2020):

$$\hat{V}(\hat{\bar{y}}_{DR}) = \hat{V}_1 + \hat{V}_2 - \hat{B}(\hat{V}_2)$$

where $\hat{V}_1 = \hat{V}\left(\frac{\sum_{i \in s_A}(\hat{y}_i)(\pi_i^R)^{-1}}{\sum_{i \in s_A}(\pi_i^R)^{-1}}\right)$ can be estimated by the usual design based estimator of the mean of the predicted values, $\hat{V}_2 = \hat{V}\left(\frac{\sum_{i \in s_B}(y_i - \hat{y}_i)(\hat{\pi}_i^B)^{-1}}{\sum_{i \in s_B}(\hat{\pi}_i^B)^{-1}}\right)$ can be estimated by $\hat{N}^{-2}\sum_{i \in S_B}\left[\frac{1-\hat{\pi}_i^B}{(\hat{\pi}_i^B)^2}\right](y_i - \hat{y}_i)^2$. $\hat{B}(\hat{V}_2)$ corrects for the bias of $\hat{V}_2$ and can be ignored in small sampling fraction settings.

- If selection probabilities π^R unknown for nonprobability sample can use Rubin's combining rule from posterior draws, using $\hat{V}(\hat{\bar{y}}_{DR})$ for known selection probabilities π^R for the within-imputation estimates of variance.

Inference Under Doubly Robust Estimation

- Alternative Bayesian method: joint model

$$\pi_i^R \mid x_i, \gamma, \phi \sim$$

$$BETA\left(\phi(exp(\gamma^T x_i)/(1+exp(\gamma^T x_i))\right), \phi/(1+exp(\gamma^T x_i))$$

$$Z_i \mid x_i, \beta \sim BER\left(exp(\beta^T x_i)/(1+exp(\beta^T x_i)\right)$$

$$Y_i \mid x_i^* \theta, \sigma \sim N(\theta^T x_i^*, \sigma^2)$$

- Obtain draws from DR estimator after drawing above parameters
- Replace parametric regressions above with relevant BART estimators and obtain draws from DR estimator.

Simulation Study

- Generate population of $A = 1,000$ cluster starting with design variables D and common covariates X:

$$\begin{pmatrix} \mathrm{D} \\ \mathrm{X}_0 \\ \mathrm{X}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \text{-}\rho/2 & \rho \\ \text{-}\rho/2 & 1 & \text{-}\rho/2 \\ \rho & \text{-}\rho/2 & 1 \end{pmatrix} \right)$$

  (Actually observe $X_2 = I(X_0 > 0)$).
- Generate continuous outcome c and binary outcome b:

$$Y^c_{ai} \mid x_a, d_a \sim N(1 + 0.5x^2_{1a} + 0.4x^3_{1a} - 0.3x_{2a} - 0.2x_{1a}x_{2a} - 0.1d_i + u_a, 1)$$

$$Y^b_{ai} \mid x_a, d_a \sim Ber\left(\frac{e^{-1+0.1x^2_{1a}+0.2x^3_{1a}-0.3x_{2a}-0.4x_{1a}x_{2a}-0.5d_i+u_a}}{1+e^{-1+0.1x^2_{1a}+0.2x^3_{1a}-0.3x_{2a}-0.4x_{1a}x_{2a}-0.5d_i+u_a}} \right)$$

- Selection probabilities for clusters:

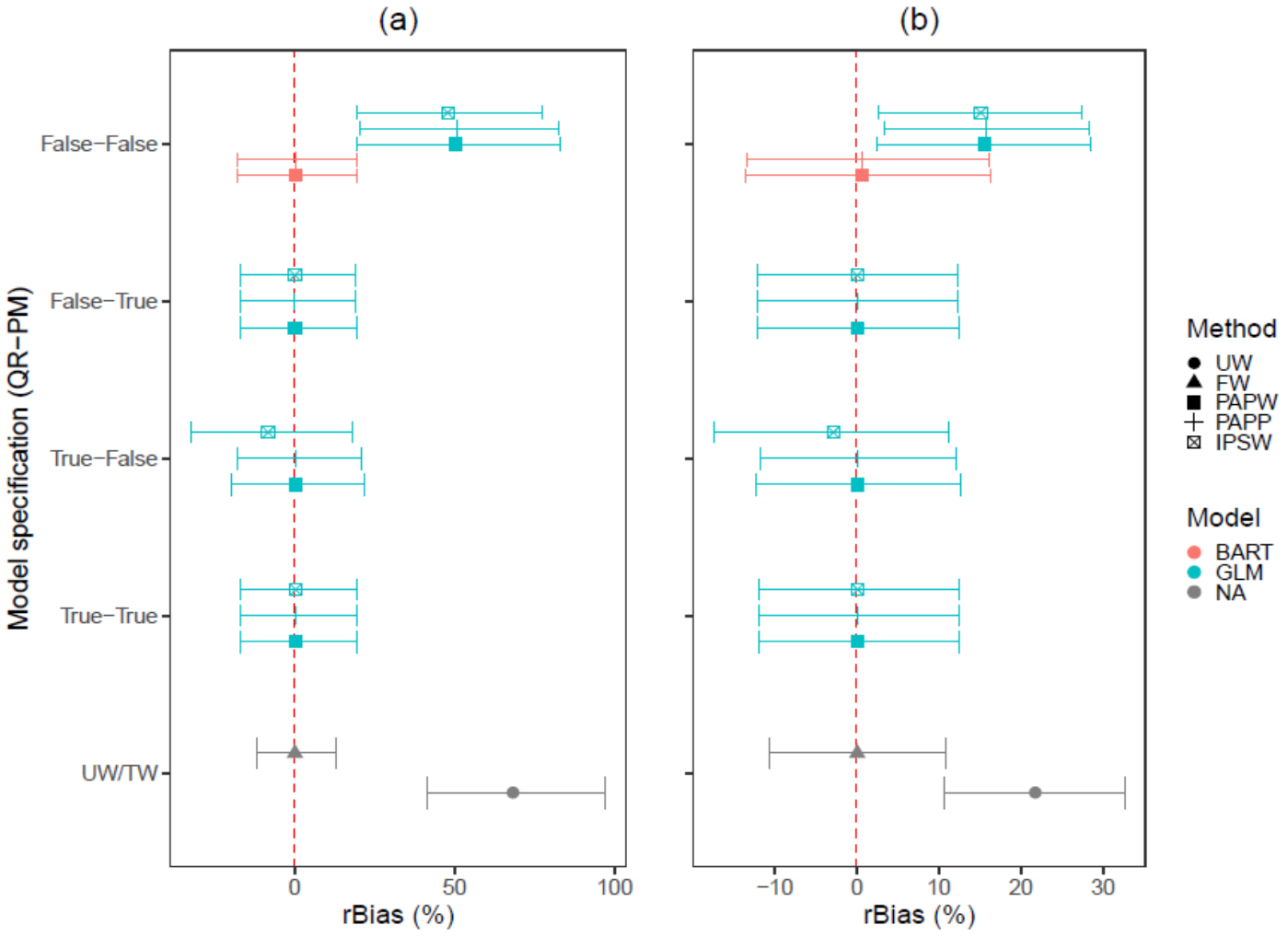
$$P(\delta^R_a = 1 \mid d_a) = \frac{e^{\gamma_0 + 0.5d_a}}{1 + e^{\gamma_0 + 0.5d_a}}$$

$$P(\delta^B_a = 1 \mid x_a) = \frac{e^{\gamma_1 - 0.1x_{1a} + 0.2x^2_{1a} + 0.3x_{2a} - 0.4x_{1a}x_{2a}}}{1 + e^{\gamma_1 - 0.1x_{1i} + 0.2x^2_{1i} + 0.3x_{2i} - 0.4x_{1i}x_{2i}}}$$

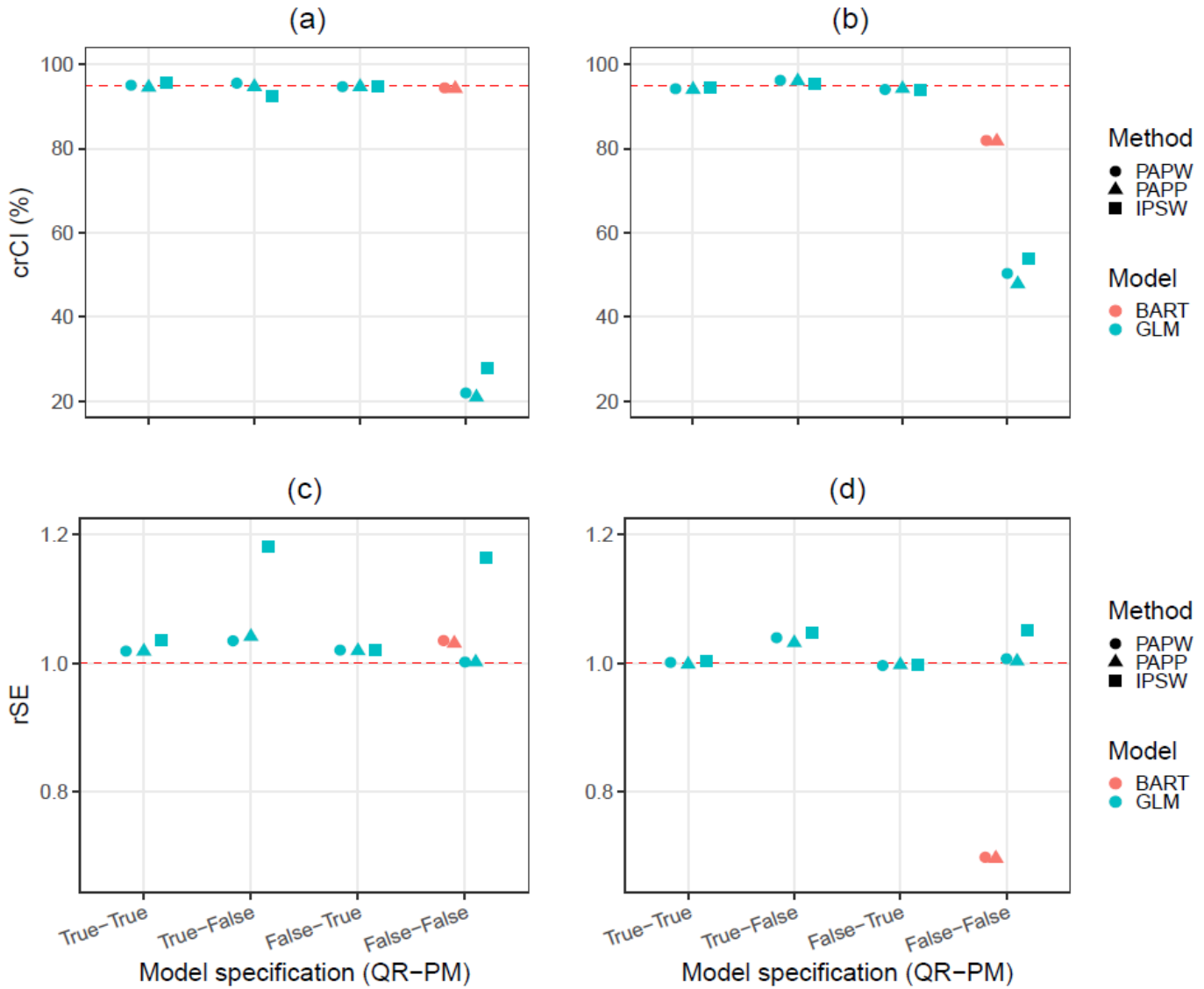
Simulation Study

- Given X Sample design is ignorable in S_B, not in S_R.
- Assume δ_i^R known only in the probability sample, δ_i^B unknown.
- Probability sample $n_R = 100$, non-probability sample $n_B = 10{,}000$, where γ_0 and γ_1 are chosen to meet these values.
- Assume $\rho = 0.2$.
- Consider two alternatives for pseudo-weights:
  - PAPW: Situation where design weights are known for non-probability sample
  - PAPP: Situation where design weights must be estimated for non-probability sample
- Also consider DR version of Chen et al. (2020) that uses their QR weight estimator (IPSW)
- Consider model misspecification by dropping interaction terms.

Simulation Study: Bias (a=continuous, b=binary)



Simulation Study: Coverage/RMSE (a,c=continuous, b,d=binary)



Summary

- This is not a complete survey of relevant literature: McConville et al. (2017) and Chen et al. (2018, 2019) consider alternative model assisted approaches that develop calibration weights to model-based estimators.
- The proposed methods work well in simulated settings but are imperfect in practice
- Lacking good covariates for assessing differences between the probability and non-probability sample together with selection being dependent on outcomes after adjustment for covariates/interaction between mean models and sample selection prevented full correction of selection bias.
- There is a need for high quality probabilty samples to collect relevant data elements for adjustment across the medical, health, and social spectrum for use in adjustment.

Open Issues and Areas for Future Research

- Combining multiple data sources (e.g. administrative records, probability sample, non-probability samples).
- Extending these methods to allow for modeling rather than just descriptive statistics: regression, small area estimation, causal inference, etc.
  - QR approach can borrow from survey literature, but DR approaches seem to require different thinking.
- Developing methods for sensitivity analyses to deal with failure of assumptions.
  - Some work has been done to address failure of ignorability by borrowing from the pattern-mixture model work in the missing data literature (Andridge 2024).
  - Extending this to the modeling setting is another open issue.

References

Andridge, R. R. (2024). Using proxy pattern-mixture models to explain bias in estimates of COVID-19 vaccine uptake from two large surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society*, in press.

Chen, J. K. T., Valliant, R. L., Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44:117-145.

Chen, J. K. T., Valliant, R. L., Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society*, C68:657-681.

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115:, 2011-2021.

References

Chipman, H. A., George, E. I., McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4.

Elliott, M.R., Resler, A., Flannagan, C., Rupp, J. (2010). Combining data from probability and non-probability samples using pseudo-weights. *Accident Analysis and Prevention*, 42:530–539.

Elliott, M.R., Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys. *Journal of the Royal Statistical Society*, C54:595–609.

References

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31:799–815.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77:237–250.

McConville, K. S., Breidt, F. J., Lee, T., Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5:131-158.

Rafei, A, Flannagan, C.A.C., Elliott, M.R. (2020). Big Data for Finite Population Inference: Applying Quasi-random Approaches to Naturalistic Driving Data using Bayesian Additive Regression Trees. *Journal of Survey Statistics and Methodology*, 8:148-180

References

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Smith, T. M. F. (1983). On the validity of inferences from nonrandom samples. *Journal of the Royal Statistical Society*, A146:394–403.

Tan, Y. V., Flannagan, C. A., Elliott, M. R. (2019). 'Robust-Squared' imputation models using BART. of Survey Statistics and Methodology, 7:465-497.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B58:267-288.

References

Van Der Laan, M. J., Polley, E. C., Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6.

Valliant, R., Dever, J. A., Kreuter, F. (2018). Nonprobability sampling. *Practical tools for designing and weighting survey samples*, 565-603.