Using Synergies between Survey Statistics and Causal Inference to Improve Transportability of Controlled Randomized Trials

Michael Elliott[1,2], Orlagh Carrol[3], Richard Grieve[4], James Carpenter[3]

[1]Department of Biostatistics, University of Michigan
[2]Survey Methodology Program, Institute for Social Research
[3]Department of Medical Statistics, London School of Hygiene and Tropical Medicine
[4]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine

Pipette to Patient to Population

- In the clinical trial world we discuss “bench to bedside” (or “pipette to patient”), bringing the results of biological research to improve patient health.
- But a missing piece is in step from the patient to the population.
- Randomized clinical trials (RCT) are the gold standard for assessing causal effects, since randomization eliminates *confounding* due to either observed or unobserved covariates (Fisher 1926).
- Randomization does *not* eliminate the effect of *effect modifiers*, which can impact the causal effect of treatment in a population that differs from the RCT sample. (Elliott 2016).
- “Transporting” the sample ACE estimators to the population ACE requires understanding the relationship between the treatment effect in the sample and the treatment effect in the population.

Casual Inference: Review and Notation

- (Population) Average casual treatment effect comparing treatment level $Z = z$ to $Z = z'$: $PATE = N^{-1} \sum_{i=1}^{N} (Y(z)_i - Y(z')_i)$. where $Y(z)$ is the "potential outcome" for the same subject Y_i under different treatment levels $Z = 0, ... T$ (Holland 1986).
- If Z is randomized and the sample=population, then observed mean $\overline{Y}_Z$ is unbiased for the PATE: $E(\overline{Y}_Z) = \overline{Y(Z)}$.
- However, if our trial is only a sample of the population, then $\overline{y}_Z$ can remain unbiased for $\overline{Y(Z)}$ only if the sampling indicator I_i is independent of $Y(Z)_i$.
  - Holds if I_i is a random sample from the population, or more generally, $P(I_i) = \pi_i$ for known π_i (replace $\overline{y}_Z$ with a weighted mean $w_i = \pi_i^{-1}$).
  - Otherwise if I_i is guided by an unknown mechanism, and $Cov(I_i, X_i) \neq 0$ for X_i where <u>interaction</u> is present $(E(Y(1)_i - E(Y(0)_i \mid X_i = x) \neq E(Y(1)_i - E(Y(0)_i \mid X_i = x'))$, $\overline{y}_Z$ may be biased for $\overline{Y(Z)}$.

Generalizability Review

- Cole and Stuart (2010) combined data from a RCT of HIV testing the effect of a protease inhibitor with data from US-wide surveillance of new HIV cases to develop inverse probability of selection weights.
- Stuart et al. (2011) developed a propensity matching method based on the propensity to be in population.
- “Doubly-robust” methods that combine propensity score weights and outcome models have been the focus of recent developments.
    - Dahabreh et al. (2020) consider three versions of these estimators that combine predictions of the outcome under treatment or control in the representative sample with IPTW-weighted residuals of the outcome model in the RCT.
    - Schmid et al. (2022) consider a targeted maximum likelihood estimator (TMLE) that uses the IPTW weight itself together with the outcome model to predict the outcome under treatment and control in the representative sample.
- Degtiar and Rose (2023) provide a overview of the currents methods used for RCT generalization.

Non-probability Inference Review

- Valliant and Dever 2011 develop IPWTs to estimate a “true” probability of selection for the non-probability sample elements in a manner similar to Cole and Stuart. Elliott et al. (2010) developes IPWTs in a somewhat differently.
- Rivers (2006) matched subjects in the non-probability sample to subjects in the probability sample via propensity to be in the probability sample, with the matched nonprobability sample used for inference.
- Direct outcome regression models that predict outcomes based on covariates are less common in the non-probability literature.
    - But “doubly robust” estimators have been developed: Chen et al. (2020) use estimators that combine model-based estimates from the probability sample with propensity-weighted residuals from non-probability sample.
- A review of estimation from non-probability samples is available at Wu (2022).

Distinctions between the Generalizability and Probability/Non-probability Sampling Literature

There are many similarities between the RCT generalizability literature and the combining of probability and non-probability samples literature, but there are also key distinctions.

- With the exception of Ackerman (2021), the generalizability literature has generally ignored complex sample design features such as weighting, clustering, or stratification in the benchmark probability sample, although these features are commonly present in both general population surveys.
- While the probability survey literature has a large section devoted to missing data, it usually does not face a setting where all observations have missing elements in a joint distribution of interest.
- The relevant patient population may be more difficult to define, let alone obtain a high-quality sample from.

Our Proposed Work: Notation and Assumptions

- Notation:
  - Defined population of size N.
  - Binary treatment $Z_i \in \{0, 1\}$, with potential outcomes $Y(0)_i$ and $Y(1)_i$.
  - Sampling indicators S_i^R (R=randomized trial) and S_i^B (B=probability/benchmark dataset)..
  - Probabilty of being sampled in B is known: $P(S_i^B = 1) = \pi_i^B$.
  - Common covariates X_i in B and R.
- Assumptions:
  - Randomization: $(Y(1)_i, Y(0)_i) \perp Z_i \mid S_i^R = 1$;
  - Stable Unit Value Treatment Assignment (SUTVA): the observed outcome $Y_i = z_i Y(1)_i + (1 - z_i) Y(0)_i$ for treatment assignment $Z_i = z_i$;
  - Positivity: $P(S_i^R = 1) > 0$ and $P(S_i^B = 1) > 0$ for all i;
  - Estimability: $P(S_i^R = 1) = \pi_i^R = g(X_i; \theta)$ for known g and unknown θ;
  - Ignorability: $(Y(1)_i, Y(0)_i) \perp S_i^R, S_i^B \mid X_i$.

Pseudo-weights

- Standard inverse probability weighting (Valliant and Dever (2011)):

$$\pi_i^{RB} = P\left(S_i^R = 1 | X_i = x_i, S_i^B = 1 \text{ or } S_i^R = 1\right).$$

  where π_i^{RB} is estimated by (weighted) logistic regression,

- Elliott et al. (2011) show via Bayes' rule that

$$\pi_i^R = P(S_i^R = 1 | X_i = x_i) \propto$$

$$P(S_i^B = 1 | X_i = x_i) \frac{P\left(S_i^R = 1 | X_i = x_i, S_i^B = 1 \text{ or } S_i^R = 1\right)}{1 - P\left(S_i^R = 1 | X_i = x_i, S_i^B = 1 \text{ or } S_i^R = 1\right)} = \pi_i^B \times \frac{\pi_i^{RB}}{1 - \pi_i^{RB}}$$

  The components of π_i^R can be estimated using generalized linear regression or Bayesian Additive Regression Trees (Chipman et al. 2010).

- Chen et al. (2020) argue that $\hat{\pi}_i^{RB}$ is not a consistent estimator of π_i^R unless π_i^R is a constant. Chen et al. suggest an maximum likelihood estimator of π_i^R that does provide a consistent estimator; however, it does not easily admit non-linear estimators such as BART.

Prediction

Under randomization, we have

$$E(Y(1)_i \mid X_i) = E(Y_i \mid X_i, Z_i = 1) \mid S_i^R = 1$$

$$E(Y(0)_i \mid X_i) = E(Y_i \mid X_i, Z_i = 0) \mid S_i^R = 1.$$

Thus a correct model of $E(Y_i \mid X_i, Z_i)$ allows prediction of $Y_i(1 - Z_i)$, and the following estimators of the PATE are

$$\hat{\Delta}_{WVD} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{RB}[Z_i(y_i - \hat{Y}(0)_i) + (1 - Z_i)(\hat{Y}(1)_i - y_i)]}{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{RB}}$$

$$\hat{\Delta}_{WE} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{R}[Z_i(y_i - \hat{Y}(0)_i) + (1 - Z_i)(\hat{Y}(1)_i - y_i)]}{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{R}}$$

Prediction

If outcome information is available in the probability sample, an alternative that only uses prediction is

$$\Delta_{PRED} = \hat{\hat{Y}}(1) - \hat{\hat{Y}}(0),$$

$$\hat{\hat{Y}}(Z) = \frac{\sum_{i=1}^{N} [I(S_i^R = 1) + (w_i^B - n_R/n_B) I(S_i^B = 1)][I(Z = z_i) y_i + I(Z = 1 - z_i) \hat{Y}_i(Z)]}{n_R + \sum_{i=1}^{N} I(S_i^B = 1)(w_i^B - n_R/n_B)}$$

Inference

- Since all of the methods we are consider involve estimating $Y(1-z_i)$ using BART, we will use a Bayesian approach for inference.
- Each draw of $Y(1-z_i)$ generates a draw of the relevant PATE estimator.
  - Point estimates are obtained as the posterior mean of these draws, with 1-α credible intervals obtained from the $\alpha/2$ and $1-\alpha/2$ empirical CDFs.
  - For Δ_{WE} we also consider an estimator of the variance (Δ_{WE2}) that incorporates uncertainty in the estimation of π_i^R.

Inference

- Because the prediction model uses a complex sample design for the probability sample, we use Rubin's Rules for combining multiple imputations:

$$\hat{E}(\Delta_{PRED} \mid \text{data}) = \frac{1}{B}\sum_{b=1}^{B} \Delta_{PRED}^{(b)}$$

$$v(\Delta_{PRED} \mid \text{data}) = \frac{1}{B}\sum_{b=1}^{B} v(\Delta_{PRED}^{(b)}) +$$

$$\frac{B+1}{B}\frac{1}{B-1}\sum_{b=1}^{B}\left(\Delta_{PRED}^{(b)} - \hat{E}(\Delta_{PRED} \mid \text{data})\right)^2$$

where $v(\Delta_{PATE}^{(b)})$ is estimated using a design-based estimator of variance that treats the imputed values of $Y(1-z_i)$ as observed.

Treatment effect among the treated

Simulations and example focus on population treatment effect among the treated (PATT):

$$\hat{\Delta}_{WVD,PATT} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{RB} Z_i(y_i - \hat{Y}(0)_i)}{\sum_{i=1}^{N} I(S_i^R = 1)Z_i/\hat{\pi}_i^{RB}}$$

$$\hat{\Delta}_{WE,PATT} = \frac{\sum_{i=1}^{N} I(S_i^R = 1)/\hat{\pi}_i^{R} Z_i(y_i - \hat{Y}(0)_i)}{\sum_{i=1}^{N} I(S_i^R = 1)Z_i/\hat{\pi}_i^{R}}$$

$$\hat{\Delta}_{PRED,PATT} = \frac{\sum_{i=1}^{N}[I(S_i^R = 1) + (w_i^B - n_R/n_B)I(S_i^B = 1)]Z_i[y_i - \hat{Y}_i(0)]}{n_{R_1} + \sum_{i=1}^{N} I(S_i^B = 1)Z_i(w_i^B - n_R/n_B)}$$

where n_{R_1} is the number of observations assigned to treatment in the RCT.

- Inference using BART for prediction proceeds as the in estimation of the PATE.

Simulation Study

- Potential outcome $Y(Z)$ for a binary treatment Z and two normally distributed covariates, X_1 and X_2:

$$Y(Z) \sim \mathcal{N}(\mu_Z, 1)$$

$$\mu_1 = \beta_0 + \delta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2, \mu_0 = \beta_0$$

- Poisson sampling is used to allocate observation i into the RCT (R; non-probability) data or benchmark (B; probability) sample ($N = 20,000$, $n = 1000$):

$$\Pr(S_i^B = 1) = \text{expit}(\psi_0^B)$$

$$\Pr(S_i^R = 1 \mid S_i^B) = \begin{cases} 0, & S_i^B = 1 \\ \text{expit}(\psi_0^R + \psi_1^R X_{1,i} + \psi_2^R X_{2,i} + \psi_3^R X_{1,i} X_{2,i}), & S_i^B = 0 \end{cases}$$

- Consider a $2 \times 2 \times 3$ design:
  - Outcome with (correctly specified) and without (misspecified) quadratic term
  - RCT SRS ($\psi_1^R = \psi_2^R = \psi_3^R = 0$), without interaction ($\psi_3^R = 0$), with interaction
  - Alignment + (Effect of X in same direction for outcome and selection) and - (different directions) (Kern et al. 2016).

Simulation Study: Summary

- SATT good if RCT is simple random sample; poor otherwise.
- WVD (estimated with logistic regression and no interaction) not too bad for bias until prediction model is complex; coverage is poor is prediction model is misspecified.
- WE1 (treating pseudo-weight as fixed) generally works reasonable well with respect to bias but has modest undercoverage with more variable selection probabilities; WE2 (incorporating variance of pseudo-weight) somewhat overcorrects for more conservative coverage except when prediction is complex, in which case bias effects coverage.
- PRED has best bias properties and, because it utilizes predictions from benchmark data, much smaller RMSE. Generally good coverage though some undercoverage occurs when prediction model is simple and positively aligned.

Study of pulmonary artery catheterization (PAC) in critical care

- PAC is an invasive and controversial cardiac monitoring device that is used in critical care. “PAC-Man” randomized trial (Harvey et al. 2005):
  - 1,013 subjects at 65 United Kingdom intensive care units.
  - Outcome=in-hospital mortality.
- Concerns about differences between the study sites and the general population in which PAC is used (Sakr et al. 2005).
- Obtain data from the Intensive Care National Audit Research Centre (ICNAR) database (Harrison et al. 2004)
  - 1.5 million admissions to 250 critical care units in the UK.
- Restricting to same inclusion and exclusion criteria as PAC-Man yields 1052 PAC population cases
- Population control group not exchangeable with RCT controls, even conditional on available covariates.
  - Restricted their analysis to the treated only: PATT
  - Approximate as being a SRS from a superpopulation by assigning a small sampling fraction value: 0.01 so $\pi_i^B \equiv 0.01$ and thus $w_i^B \equiv 100$.

Covariates

Variable	RCT	INCAR	*p*-value
Age	64.5	61.9	<0.001
% Female	41.8	39.0	0.22
% Elective	6.3	9.3	0.007
% Emergency	27.4	23.1	
% Medical	66.2	67.6	
% Ventilator	90.3	86.2	0.006
% Teaching Hosp.	21.5	41.2	<0.001
Survival Prob.	54.1	52.5	0.15
AP2 score	17.8	17.5	0.32
% Cardio event	3.8	3.2	0.60
% Renal failure	1.2	1.2	1.00
% Resp problems	3.6	2.5	0.19
% Liver failure	2.5	2.2	0.78
% Immunte disorder	7.8	6.8	0.46
Glasgow coma score	3.95	3.77	0.042

Results

- Adjusted SATT obtained from a BART model trained on the observed data in the RCT assigned to treatment assigned to control:-4.3% (95% CI -9.5%,1.0%)
- The PATT estimated under the pseudo-weighting method of WE1 is 0.2% with a 95% CI of (-4.2%,4.4%).
- The PATT estimated under WE2 is 0.2% with a 95% CI of (-9.5%,10.1%).
- The PATT estimated under PRED was 6.8% with a 95% CI of (-1.2%,14.8%).
    - While none of the effects significant, the PATT expected direction of the effect, in contrast to the SATT.

Model Checking: Testing for ignorability

- Transportability relies on the ignorabilty assumption: potential outcomes are independent of sampling indicator given covariates.
  - Impute $Y(Z)_i$ when $Z_i = 1 - z$ in the probability sample (or $Y(Z)_i$ for $Z = 0, 1$ if Y is not observed in the probability sample).
- Assumption testable when Y is observed in the probability sample
  - Test the reduced version $Y(1)_i \perp S_i^R, S_i^B \mid X_i$ in PAC-Man since only treatment outcomes are available.
- Posterior predictive distribution p-value: $T^{rep} = \sum_{i=1}^{N} I(S_i = 1) Y(1)_i^{rep}$ versus $T^{obs} = \sum_{i=1}^{N} I(S_i = 1) I(Z_i = 1) y_i$.
  - $P(T^{rep} < T^{obs} \mid \text{ data }) = 0.159$,
  - Overestimate the success of PAC in the population, or, equivalently, subjects in the RCT were more likely to have a good outcome even after controlling for available covariates.

Ignorability-corrected PATT

- The impact on the failure of ignorability in this setting depends on how the joint distribution of $(Y(1)_i, Y(0)_i) \mid X_i, S_i = 2$ differs from $(Y(1)_i, Y(0)_i) \mid X_i, S_i = 1$.
- If $\delta(1, X_i)^S - \delta(0, X_i)^S = 0$ for all X_i, $\delta(z, X_i)^S = P(Y(z)_i \mid X_i, S_i = 2) - P(Y(z)_i \mid X_i, S_i = 1)$ then the PATT estimate remains unbiased
- Other extreme:ignorability holds on the control arm, so that $\delta(0, X_i)^S = 0$ or, more generally $E(\delta(0, X_i)^S) = 0$.
  - $E(\delta(1, X_i)^S) = E(T^{rep} - T^{obs} \mid \text{ data }) = 8.4\%$
  - Estimate corrected PATT of 6.8%+8.4%=15.1%.

Discussion

- Econometricians, epidemiologists, and biostatisticians have independently invented and reinvented the wheel of causal inference for the past several decades, in the process following or borrowing the tools of population inference from survey statistics.
- Survey statistics can return the favor by adapting recently developed methods for non-probability samples for the important task of transporting randomized trials to better understand how novel treatments can work in a larger population.
- “Tip of the iceberg” of research opportunities:
  - Accommodating non-compliance.
  - Mediation; confounding by indication in longitudinal studies.
  - Adaptive trial design to ferret out key interactions.

References

Ackerman,B., Lesko, C.R., Siddique, J., et al. Generalizing randomized trial findings to a target population using complex survey population data. *Statistics in Medicine*, 40, 1101-1120.

Baker, R., Brick, J.M., Bates, N.A. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodolology* 2013, 1, 90-143.

Chipman, H.A., George, E.I., McCulloch, R.E. (2010) BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4, 266-298.

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

References

Cole, S.R., Stuart, E.A. (2010) Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Americal Journal of Epidemiology*, 172, 107-115.

Dahabreh, I.J., Robertson, S.E., Steingrimsson, J.A., et al. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39, 1999-2014.

Degtiar, I., Rose, S. (2023) A Review of Generalizability and Transportability. *Annual Review of Statistics and Its Application*, 10, 7.1-7.24.

Elliott, M.R. (2016). Discussion of ‘Perils and Potentials of Self-Selected Entry to Epidemiological Studies and Surveys’. *Journal of the Royal Statistical Society A: Statistics in Society*, 179, 357.

References

Elliott, M.R., Resler, A., Flannagan, C.A., Rupp, J.D. (2010). Appropriate analysis of CIREN data: using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Fisher, R.A. (1926) The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 503-513.

Harrison D.A., Brady A.R., Rowan K. (2004) Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit Research Centre Case Mix Programme Database. *Critical Care*, 8, 1-3.

References

Harvey S., Harrison D.A., Singer M., et al. (2005) Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomised controlled trial. *The Lancet*, 366, 472-477.

Hartman, E., Grieve, R., Ramsahai, R., et al. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society A*, 178, 757-778.

Kern, H., Stuart, E.A., Hill, J., Green, D.P. (2016) Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9, 103-127.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *J R Stat Soc*, 97, 558–625.

References

Sakr Y., Vincent J.L., Reinhart K., et al. (2005). Sepsis Occurrence in Acutely Ill Patients Investigators. Use of the pulmonary artery catheter is not associated with worse outcome in the ICU. *Chest*, 128, 2722-2731.

Schmid, I., Rudolph, K.E., Nguyen, T.Q., et al. (2022). Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations. *Communication in Statistics: Simululation and Computation*, 51, 4326-4348.

Stuart, E.A., Ackerman, B., Westreich, D. (2018) Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on Social Work Practice*, 28, 532-537.

References

Stuart, E. A., Cole, S. R., Bradshaw, C. P., Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 369-386.

Valliant, R., Dever, J.A. (2011) Estimating propensity adjustments for volunteer web surveys. *Sociological Methods Research*, 40, 105-137.

Wu, C. Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 283-311.