Data Integration in Surveys & Clinical Trials

Yajuan Si and Michael Elliott

Survey Research Center, Institute for Social Research
Department of Biostatistics, School of Public Health
University of Michigan, Ann Arbor

Population Dynamics and Health Program Workshop

September 24, 2024

Motivation

- Statistics quality metrics consider both study **design** and **analysis**.
- A single data source often cannot meet the growing demand for high-quality statistical products.
- In this "Big Data" era, it is crucial to understanding how to use a variety of data sources in a given analysis.
- Key question: *Can we make valid inference from the integrated data sources for the target population?*

Combining information from multiple sources

1. Multiple surveys: web-based and face-to-face instruments; main study and refreshment samples
2. Surveys and administrative data: e.g., National Experimental Well-Being Statistics at the Census Bureau
3. Survey and commercial data: e.g., the Panel Study of Income Dynamics and Zillow housing data
4. Survey and Medicare data: e.g., the Health Retirement Study and data from the Centers for Medicare & Medicaid Services
5. Probability and nonprobability surveys: e.g., the Behavioral Risk Factor Surveillance System and the All of US Study
6. And more others (summary information from historical studies, auxiliary margins, meta-analysis, etc.)

Assessing data quality of an integrated dataset

1. **Relevance**: Whether the data are responsive to users' analytic needs in terms of subject content, population covered, adequacy for estimation, frequency, and utility for longitudinal as well as repeated cross-sectional analysis;
2. **Timeliness and punctuality**: How long the interval is between the phenomena, when data are available, and whether they are produced punctually in accordance with an announced publication schedule;
3. **Accessibility and clarity**: Whether users can easily obtain and understand the data and associated documentation;
4. **Coherence and comparability**: Whether the data accord with standard definitions and are consistent over time and with other relevant data; and
5. **Granularity**: The extent to which the data can be provided for population groups and geographic areas.

Smeeding et al. (2024)

Combining scenarios

- Combining summary statistics
  - Dual frame estimation
- Combining microdata
  - Overlap of some individuals across data sources
    - Record linkage (deterministic and probabilistic)
  - None or unknown overlap across data sources
    - Statistical Matching/Data Fusion
    - Probability/Nonprobability sampling

Probability survey

1. Defined (finite) population
2. Sampling frame that contains all elements of the population and only those elements
3. Every element in the population has a non-zero probability of selection
   - Not necessarily equal – may purposely over-/under-sample some parts of population
4. Can define (in theory) all possible samples that can be obtained from the sampling procedure
   - Design features: **strata** to improve statistical efficiency, **clusters** to ease logistics/improve cost efficiency, and **weights** as the inverse of a known probability of selection is associated with each sample
5. The chosen sample is selected via a randomized mechanism from all possible samples

Dual frame estimation

- May have several available data sources to estimate a measure of interest.
- Often have multiple overlapping sampling frames, e.g.,
  - Face to face and web-based online panels
  - Cell phones and landline phones
- Assume variables are measured identically.

Dual frame estimation cont.

If we know how population is distributed across frames, Hartley (1962) showed a minimum variance estimator of a total is given by

$$\hat{Y} = N_a \bar{y}_a^A + N_b \bar{y}_b^B + N_{ab} \left(p \bar{y}_{ab}^A + (1-p) \bar{y}_{ab}^B \right),$$

where

$$p = \frac{(n_a + n_{ab})/(N_a + N_{ab})}{(n_a + n_{ab})/(N_a + N_{ab}) + (n_b + n_{ab})/(N_b + N_{ab})}.$$

- N_a and n_a = population in frame/sample A only
- N_b and n_b = population in frame/sample B only
- N_{ab} and n_{ab} = population/sample in overlap
- $\bar{y}_a^A$ and $\bar{y}_b^B$ are the mean estimates in sample A and B only
- $\bar{y}_{ab}^A$ and $\bar{y}_{ab}^B$ are the mean estimates for all of A and all of B (including overlap) respectively.

Dual frame estimation cont.

If we know N_a and N_b but not the overlap size of frames, Fuller and Burmeister (1972) gave a minimum variance total estimator of

$$\hat{Y} = (N_A - \hat{N}_{ab})\bar{y}_a^A + (N_B - \hat{N}_{ab})\bar{y}_b^B + \hat{N}_{ab}\bar{y}_{ab},$$

where

- $\bar{y}_{ab} = \frac{n_{ab}^A \bar{y}_{ab}^A + n_{ab}^B \bar{y}_{ab}^B}{n_{ab}^A + n_{ab}^B}$
- $\hat{N}_{ab}$ is the smallest root that solves
  $(n_A + n_B)N_{ab}^2 + (n_A n_b + n_a n_B + n_{ab}^A N_A + n_{ab}^B N_B)N_{ab} +$
  $(n_{ab}^A + n_{ab}^B)N_A N_B = 0$

Skinner and Rao (1996) extended to general sample designs, replacing simple random sample estimators with design-consistent estimators.

Record linkage

- Linking records from one dataset to another.
- Typically between a survey and administrative records, or between two sets of administrative records.
- Trivial when there are high-quality unique identifiers in the dataset (Social Security or tax ID numbers, university IDs, etc.)
  - Deterministic record linkage
- Otherwise need to rely on common elements whose joint distribution is (hopefully) unique: name, address, birth date, for example.
  - Probabilistic record linkage

Probabilistic record linkage

Large literature in this area; key issues revolve around typographical errors or misspelling in text fields ("An Arbor" or "Anne Arbor"), transpositions such as day and month vs. month and day (03/09/18 vs. 09/03/18) or ambiguities due to nicknames, etc. ("Mike Elliott" vs. "Michael Elliott").

- Each potential link is assigned a probability of being correct conditional on the observed data.
- Failure to accurately link may yield biased results; subjects with more errors or ambiguity may be systematically different from other subjects.
  - Selection bias akin to missingness.

Bayesian record linkage (Steorts et al., 2016)

- Assume we have k files with p fields in common, which are all categorical with M_l levels.
- Observed data x_{ijl}=lth data field, $l = 1, ..., p$, in ith list, $i = 1, ..k$, for the jth element in the ith list.
- Let $\lambda_{ij} \in \{1, ..., N\}$ denote the true record number, $y_{\lambda_{ij}l}$ be the true value of x_{ijl}, and z_{ijl} be an indicator for whether x_{ijl} is incorrect in the observed database.

$$
\begin{aligned}
&x_{ijl}|\lambda_{ij}, y_{\lambda_{ij}l}, z_{ijl}, \theta_l \sim \begin{cases} = y_{\lambda_{ij}l} \text{ if } z_{ijl} = 0 \\ Multi(1, \theta_l) \text{ if } z_{ijl} = 1 \end{cases} \\
&z_{ijl}|\pi_l \sim Bernoulli(\pi_l) \\
&y_{j'l}|\theta_l \sim Multi(1, \theta_l) \\
&\theta_l \sim Dirichlet(1/M_l, ..., 1/M_l) \\
&\pi_l \sim UNI(0, 1) \\
&p(\lambda_{ij}) \propto 1
\end{aligned}
$$

- All these quantities (including N) are unobserved
- Fit using Bayesian Markov chain Monte Carlo (MCMC) methods

Bayesian record linkage cont.

- Probability of a linkage between two records (i, j) and (m, n) is given by the proportion of MCMC draws for which $\lambda_{ij} = \lambda_{mn}$.
- A few assumptions are required to make this model identifiable:
  - The error rate π_l for a given variable is the same regardless of the observation or the list.
  - The distribution of the true values y_{jl} is the same regardless of the observation or the list
  - All of the possible linkage structures indexed by λ are equally likely.

Probabilistic record linkage

- If the number of records in large (say more than several thousands) approaches like Steorts et al. (2016) may be impractical (quadratic in n).
  - Use "blocking" (Hernandez and Stolfo, 1998).
  - "Prematch" based on one or more observed attributes into blocks, and then conduct matching only within these blocks.
    - Risk of false negatives: subjects cannot be matches across blocks if there is error in the prematching attributes.
    - Can extend to "multiple blocking": create sets of overlapping blocks (Gravano et al., 2001).
    - Balance risk of false negatives again time cost of limited blocking.

Statistical matching/data fusion

- Despite the name, this does not involve matching or linkage of records.
- Used when few or no records are believed to occur in both datasets (when working with two independent sample surveys with small sampling fractions, for example).
- Requires a common set of covariates in the two datasets, as well as interest in understanding the association between two sets of covariates that are unique to one of the datasets.

Statistical matching/data fusion cont.

Survey	Y (Heart disease)	Z (Age, gender, etc.)	X (Nutritional intake)
A			?
B	?		

- Typically assume independent sampling between the two samples.
- Think of $(X\ Z)$ (survey B) as the "donor" dataset and $(Y\ Z)$ (survey A) as the "recipient" dataset.
- Goals is to obtain $f(X, Y, Z)$ for inference.

Statistical matching/data fusion cont.

Matching algorithm:

Find observations (x_i, z_i^B) in survey B and (y_j, z_j^A) in survey A with $z_i^B = z_j^A$, and "donate" (impute) x_i, creating (x_i, y_j, z_j^A).

- If there are multiple observations in a subset of S such that $z_{i \in S}^B = z_j^A$, chose one at random
- If z is continuous or categorical with many categories so that there are no observations with $z_i^B = z_j^A$, construct a distance measure and choose the ith observation that minimizes $d(z_i, z_j)$.

Statistical matching/data fusion cont.

- If samples A and B have the same design, then the resulting joint distribution will be given by

$$f(X, Y, Z) = f(X, Z)f(Y|Z)$$

- This implies conditional independence between X and Y given Z:

$$\begin{aligned} f(X, Y, Z) &= f(X, Z)f(Y|Z) \\ &= f(X|Z)f(Z)f(Y|Z) \\ &= f(X|Z)f(Y, Z) \end{aligned}$$

- So conditioning both sides on Z yields

$$f(X, Y|Z) = f(X|Z)f(Y|Z)$$

- Thus marginal distributions remain the same after matching.
- Joint distribution depends on how reasonable the assumption of conditional independence is.

Integrating probability and nonprobability samples

- Suppose we have a setting like statistical matching, but the qualities of Surveys A and B are different?
- Survey A, which contains the outcome of interest, is a nonprobability survey.
  - Can extend to settings where the missing Y and/or the missing X are observed $\rightarrow$ concerns about selection bias in A.
- The availability of data from a probability sample allows the possibility of using information from that sample to adjust for selection bias in a nonprobability sample.
- We will discuss two approaches for incorporating this information:
  - Quasi-randomization ("pseudo-weights")
  - Superpopulation modeling ("Mr. P")

Quasi-randomization

- Basic idea is to combine the probability sample data and nonprobabilty sample data to estimate the (pseudo-)probability of the nonprobability sample data to have been selected.
- Reciprocal of this probability can be used as a regular sample weight.
  - Not targeted to a particular variable of interest

Scenario 1: Outcome is only available in the nonprobability survey

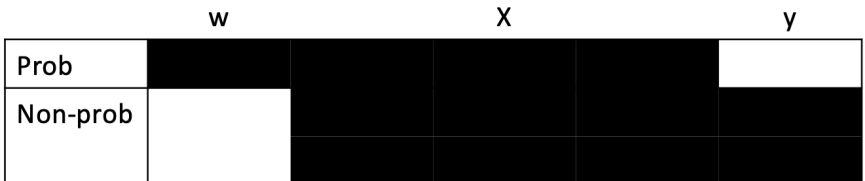


Figure: Integrating a nonprobability survey with a reference probability survey.

Setup

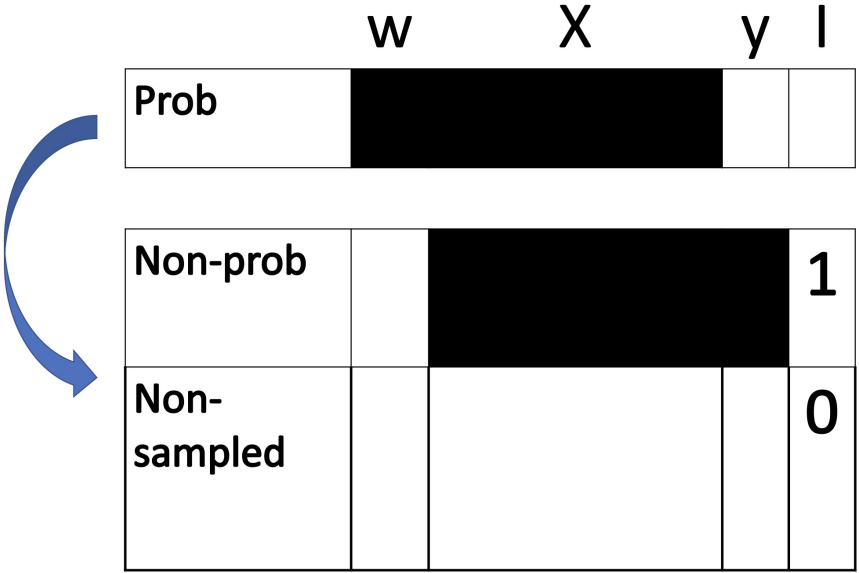
- Based on conceptual scenario 1, Elliott and Valliant (2017); Wu (2022) have reviewed existing inference approaches with nonprobability surveys.
- Suppose the target population has N units with study variable of interest y and auxiliary variables X.
- Let $\{(y_i, X_i), i \in S_{np}\}$ be the nonprobability survey data S_{np} with n_{np} participants.
- Let $I_i(= 1/0)$ be the binary inclusion indicator for unit i being included in the nonprobability sample S_{np} and defined for all population units.

Assumptions

Four basic assumptions are introduced to be able to make inference with S_{np}.

1. **Positivity**: All population units have a non-zero probability to be included in the nonprobability sample, i.e., $\Pr(I_i = 1) > 0$.
2. **Independence**: The inclusion indicators are mutually independent of each other, given the auxiliary and survey variables.
3. **Ignorability**:The indicator and the study variable are independent given the set of covariates.
4. **Common support**:There exists a probability survey sample S_p of size n_p with information on X, but not on y, in the data set $\{(w_i^p, X_i), i \in S_p\}$, where w_i^p is the sampling weight for S_p.

Leveraging the reference probability sample



Leveraging the reference probability sample

- The inferential approaches we consider use the reference probability survey to obtain the population distribution of auxiliary variables for
  - 1. estimating the model for the propensity score $\pi_i^{np} = \Pr(I_i = 1 \mid X_i)$: Weighting
  - 2. predicting the outcome of population units with the outcome model $(y_i \mid X_i)$: Imputation
  - 3. both 1) and 2): Doubly robust estimators

Inverse propensity weighting

- This quasi-randomization approach estimates the inclusion probability of the nonprobability samples: $\pi_i^{np} = \Pr(I_i = 1 \mid X_i)$.
- Under a chosen method for the estimated inclusion propensity $\hat{\pi}_i^{np}$, the inverse propensity weighted (IPW) estimator for $\bar{Y}$ is

$$\hat{Y} = \frac{\sum_{i \in S_{np}} y_i / \hat{\pi}_i^{np}}{\sum_{i \in S_{np}} 1 / \hat{\pi}_i^{np}}.$$

Pseudo-weights

- Elliott (2009) has derived the pseudo-weights as:

$$1/\pi_i^{np} = \frac{1}{\pi_i^p} \frac{\Pr(i \in S_p \mid i \in S_{np} \cup S_p,\ X_i)}{\Pr(i \in S_{np} \mid i \in S_{np} \cup S_p,\ X_i)},$$

where π_i^p is the probability of nonprobability sample units being included in the probability sample, which in principle should be known or estimated for all population units.

- The goal is to adjust the nonprobability sample weights to look like those in the probability sample.
- Rafei et al. (2020) have applied Bayesian additive regression tree models to estimate the pseudo-weights.

Weighting based on pseudo maximum likelihood

- Chen et al. (2020) estimate a maximum likelihood function $\prod_{i=1}^{N}(\pi_i^{np})^{I_i}(1-\pi_i^{np})^{(1-I_i)}$ with the pseudo log-likelihood function as

$$l = \sum_{i \in S_{np}} log\left(\frac{\pi_i^{np}}{1-\pi_i^{np}}\right) + \sum_{i \in S_p} w_i^p log(1-\pi_i^{np}).$$

- Maximum likelihood estimation can be replaced by alternatives, such as estimation equation-based methods, nonparametric methods, and tree-based methods.

Calibration weighting (Kott, 2006; Särndal, 2007)

- Poststratification balances the sample cell sizes with the population sizes in the joint contingency table to adjust the full cross-tabulation (Holt and Smith, 1979).
  - However, poststratification is often unfeasible due to small or empty sample cell sizes or unknown joint population distribution.
- Calibration selects an objective function to minimize the difference between the sample and the population (e.g., entropy balancing) and often combines with prediction models (Deville and Särndal, 1992).
- In practice, raking is the most popular calibration weighting method, which generates weights for each case to match the population margins (Si and Zhou, 2021).

Model-based predictions

- The model-based predictions use estimated parameters based on the nonprobability survey and auxiliary information about the population to generate population outcome values, $\hat{\mu}_i = E(y_i \mid X_i)$, for $i = 1, \ldots, N$.
- Flexible models or machine learning algorithms can be used as the imputation engine:
  - *big data and small probability survey*;
  - *small nonprobability survey and big census studies*.
- Similar to mass imputation (e.g., Kim et al. (2021)).

Examples

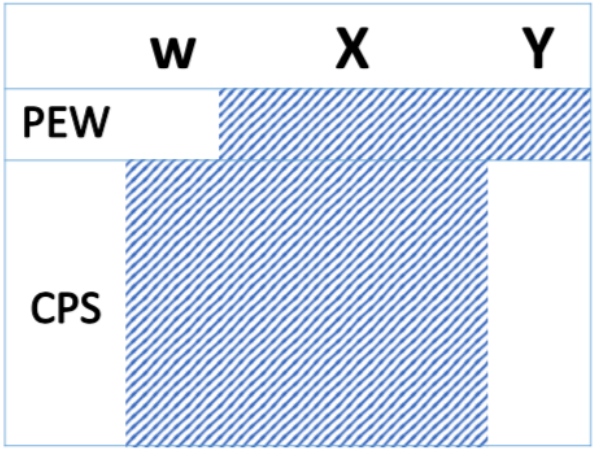


Figure: Integrate the Pew Research Organization's Political Survey (S_{np}) and the Current Population Survey (CPS, S_p).

Multilevel regression and poststratification (MRP; Gelman and Little (1997); Si (2024))

- Originally applied to estimate state-level public opinions from sociodemographic subgroups using sample surveys, MRP has two key components:
  1. multilevel regression for small area estimation by setting up a predictive model with a large number of covariates and regularizing with Bayesian prior specifications, and
  2. poststratification to adjust for selection bias and correct for imbalances in the sample composition.
- We will demonstrate the MRP interface that is publicly available.

Key to success: Highly predictive variables

- The flexible models under MRP improve small area estimation and facilitate finite population inferences, with examples including Bayesian hierarchical models with high-order interactions and global-local shrinkage prior specifications (Ghitza and Gelman, 2013; Si et al., 2020), Gaussian process regression models (Si et al., 2015), and machine learning algorithms (Broniecki et al., 2022).
- The key to success is the existence of highly predictive auxiliary variables (Si, 2024).
- Often the relevant auxiliary information is partially collected and requires modeling to generate the synthetic population distribution (Reilly et al., 2001; Kastellec et al., 2015; Kuriwaki et al., 2023; Li and Si, 2024).

Doubly robust inference

- Doubly robust (DR) estimators improve both estimators by combining IPW with a prediction model for the survey outcome and provide protection against either model misspecification (Bang and Robins, 2005).
- Chen et al. (2020) propose the DR estimator combining the probability and nonprobability surveys

$$\hat{\mu}_{DR} = \frac{1}{\sum_{i \in S_{np}} (\hat{\pi}_i^{np})^{-1}} \sum_{i \in S_{np}} \frac{y_i - \hat{\mu}_i}{\hat{\pi}_i^{np}} + \frac{1}{\sum_{i \in S_p} w_i^p} \sum_{i \in S_p} w_i^p \hat{\mu}_i.$$

Assumptions: Revisit

- In sum, strong assumptions are necessary to make inference with a nonprobability survey:
  1. The sampled units are exchangeable with nonsampled units that share the same measured characteristics,
  2. No parts of the population are systematically excluded entirely from the sample, and
  3. The composition of the sampled units with respect to observed characteristics either matches or can be adjusted to match the composition of the larger population.

Scenario 2: Outcome is available in both surveys

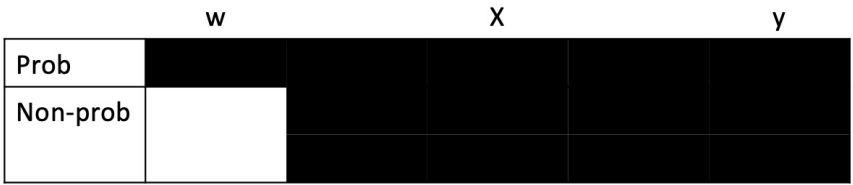


Figure: Integrating a nonprobability survey with a reference probability survey.

Scenario 2: Outcome is available in both surveys

- IPWT is similar except now outcome can be used in the estimation of the weight:
  - $\Pr(i \in S_p \mid i \in S_{np} \cup S_p,\ X_i, Y_i)$
- Alternative approach:
  - Use the nonprobability sample to construct a prior for the probability sample when estimating regression coefficients (Wiśniowski et al., 2020)
    - Incentivize the higher-quality data (S_p) by giving decreasing weight to the lower-quality "prior information" (S_{np}) as the number of high-quality observations increases
  - Nandram and Rao (2021) use power prior and weighted likelihood, where the power is based on adjusted survey weights and effective sample size.

"The choice of statistical methods is secondary to the choice of auxiliary variables" (AAPOR Task Force, 2023)

- Kang and Schafer (2007) show that the regression prediction estimator outperforms DR estimators with a predictive model.
- Empirical comparisons show that the adjustment approaches must include all predictive auxiliary variables related to the inclusion mechanism and survey variables to yield valid inference (Mercer et al., 2018; Valliant, 2020).
- Variance estimation can use linearization approximation, replication methods, and Bayesian posterior inference.

Improving transportability of a randomized controlled trial using quasi-randomization or prediction

- Randomized controlled trials (RCTs) have been the gold standard for assessing causal effects because they can eliminate both observed and unobserved confounding.
- When there are both effect modification and systematic differences between the trial sample and the ultimate population of inference with respect to these modifiers, estimates of causal effects at the population level from randomized controlled trials can still be biased.
- Transporting causal effect estimates from trials to populations requires that there is a population-representative reference sample: Integrating the RCT nonprobability sample with a probability reference sample.

Case 1

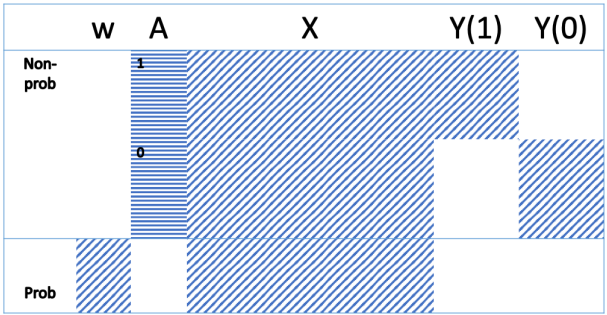


Figure: Unknown treatment assignment and outcome in the probability sample.

Case 2

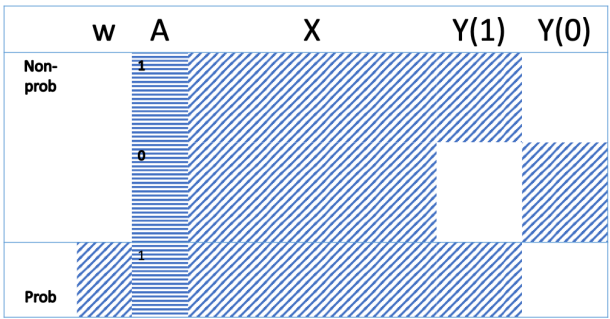


Figure: The probability sample includes cases under treatment.

Case 3

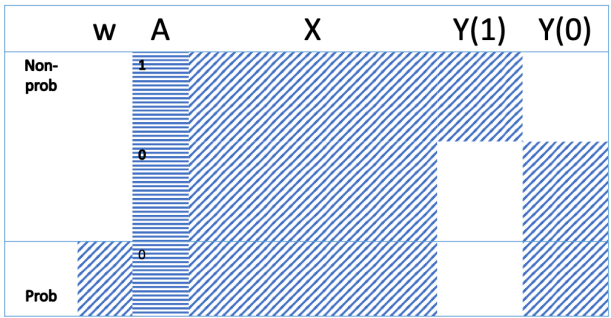


Figure: The probability sample includes cases under control.

General methods

- Weighting (Cole and Stuart, 2010)
- Matching (Stuart et al., 2011)
- Outcome regression (Kern et al., 2016)
- Doubly robust estimation (Dahabreh et al., 2020)

Degtiar and Rose (2023) have reviewed various approaches considered in the literature.

Key assumptions

- Randomization
- Stable Unit Treatment Value Assignment
- Positivity
- Common support
- Ignorability (potential outcomes have the same distribution under both data sources conditional on covariates)

Population average treatment effect on the treated (PATT)

$$\Delta_{PATT} = \frac{\sum_{i=1}^{N}(Y_i(1) - Y_i(0))I(A_i = 1)}{\sum_{i=1}^{N} I(A_i = 1)},$$

where N is the population size.

Combine RCT with a probability sample

- Inverse probability weighting (IPW)
  1. Concatenated: $w_i^c = 1/Pr(S_i^{np} = 1 \mid X_i,\, S_i^{np} = 1\, or\, U_{S_i^p=1})$
  2. Adjusted concatenated: $w_i^c - 1$
  3. Selection: $w_i^{np} = w_i^p \times \frac{Pr(S_i^p=1|X_i,\, S_i^{np}=1\, or\, S_i^p=1)}{Pr(S_i^{np}=1|X_i,\, S_i^{np}=1\, or\, S_i^p=1)}$.
- Prediction:
  1. Under randomization, assume $E(Y_i(a) \mid X_i) = E(Y_i \mid X_i, A_i = a, S_i^{np} = 1)$
  2. Under ignorability, replace $Y_i(1) - Y_i(0)$ with $y_i - \hat{Y}_i(0)$ for treated subjects and $\hat{Y}_i(1) - y_i$ for subjects assigned to the control group.
  3. Use Bayesian Additive Regression Trees (BART) for robust prediction of potential outcomes (Chipman et al., 2020)

PATT estimators (Elliott et al., 2023)

1. Combine IPW with prediction. With w_i^{np} for the RCT

$$\Delta ipw_{PATT} = \frac{\sum_{i=1}^{N} w_i^{np} I(S_i^{np} = 1) I(A_i = 1)(y_i - \hat{Y}_i(0))}{\sum_{i=1}^{N} w_i^{np} I(S_i^{np} = 1) I(A_i = 1)} \quad (1)$$

2. Impute potential outcomes of RCT and the probability sample

$$\Delta_{PATT}^{pred} = \frac{\sum_{i=1}^{N} [I(S_i^{np} = 1) + (w_i^p - n_{np}/n_p) I(S_i^p = 1)] I(A_i = 1)(y_i - \hat{Y}_i(0))}{\sum_{i=1}^{N} [I(S_i^{np} = 1) I(A_i = 1) + (w_i^p - n_{np}/n_p) I(S_i^p = 1) I(A_i = 1)]} \quad (2)$$

3. Make inference with Bayesian posterior samples or multiple imputation combining rules

MRP interface (Si et al., 2024)

- Local installation:
  `https://github.com/mrp-interface/shinymrp`
  1. install app from GitHub:
     remotes::install_github('mrp-interface/shinymrp')
  2. launch the app: *shinymrp::run_app()*
- Web demonstration:
  `https://mrpinterface.shinyapps.io/shinymrp`

Example 1: Public opinion forecasting

- The example data from the 2018 Cooperative Congressional Election Study include a dichotomous outcome measure, *Allow employers to decline coverage of abortions in insurance plans (Support / Oppose)*, and geographic-demographic predictors as below:
  - State: 50 US states
  - Age: 18-29, 30-39, 40-49, 50-59, 60-69, 70+
  - Gender: Female, Male
  - Race/ethnicity: (Non-Hispanic) White, Black, Hispanic, Other
  - Education: No HS, HS, Some college, 4-year college, Post-grad.
- MRP has two key steps: (1) fit a multilevel model for the response with the adjustment variables based on the input data; and (2) poststratify using the population distribution of the adjustment variables.

Model

- We consider a logistic regression and include varying intercepts for age, race/ethnicity, education, and state, where the variation for the state-varying intercepts is explained by the state-level predictors.

$$\Pr(y_i = 1) = \text{logit}^{-1}(\alpha^{\text{state}}_{\text{s[i]}} + \alpha^{\text{age}}_{\text{a[i]}} + \alpha^{\text{eth}}_{\text{r[i]}} + \alpha^{\text{educ}}_{\text{e[i]}} + \beta^{\text{sex}} \cdot \text{sex}_{\text{i}}) \tag{3}$$

where

- $\alpha^{\text{age}}_{\text{a}}$ is the effect of subject i's age on the tendency of having a positive response.
- $\alpha^{\text{eth}}_{\text{r}}$ is the effect of subject i's race/ethnicity on the tendency of having a positive response.
- $\alpha^{\text{educ}}_{\text{e}}$ is the effect of subject i's education on the tendency of having a positive response.

Model cont.

- $\alpha_{\mathrm{s}}^{\mathrm{state}}$: The effect of subject i's state on the tendency of having a positive response. As we have state-level predictors, we need to build another model in which $\alpha_{\mathrm{s}}^{\mathrm{state}}$ is the outcome of a linear regression with state-level predictors. For state s,

$$\alpha_{\mathrm{s}}^{\mathrm{state}} = \vec{\alpha}\vec{Z}_s^{\mathrm{state}} + e_s,$$

  with e_s as the random error.

- In the Bayesian framework we assign hierarchical priors to varying intercepts α^{name} or error terms e_s:

$$\alpha^{\mathrm{name}} \sim normal(0, \sigma^{\mathrm{name}}),\ \sigma^{\mathrm{name}} \sim normal_+(a, b),$$

  for $name \in \{\mathrm{age}, \mathrm{race}\}$. Here, $normal_+(a, b)$ represents a half-normal distribution with the mean a and standard deviation b restricted to positive values, with pre-specified values of (a, b).

MRP estimator

- To generalize results from this model to a national or subgroup estimate, we obtain the population cells in the contigency table of sex, age, race/ethnicity, education, and state and weight the model predictions by the population frequency of the poststratification cells N_j's.
- Suppose the cell-wise estimate based on model (3) is θ_j in cell j, the MRP estimate can be expressed as:

$$\theta^{MRP} = \frac{\sum N_j \theta_j}{\sum N_j}.$$

- Small area estimation is one of the main applications of MRP, and the MRP estimator for state s is

$$\theta_s^{MRP} = \frac{\sum_{j \in s} N_j \theta_j}{\sum_{j \in s} N_j}.$$

Example 2: COVID-19 infection tracking

- In the absence of comprehensive or random testing, we have developed a proxy method for synthetic random sampling to estimate the actual viral incidence in the community, based on viral RNA testing of asymptomatic patients who present for elective procedures within a hospital system.
- We collect routine testing data on SARS-CoV-2 exposure among outpatients and perform statistical adjustments of sample representation using MRP (Covello et al., 2021; Si et al., 2022).

Synthetic random sampling with hospital data

- Elective patients for invasive procedures in the hospital with a lack of symptoms and a negative exposure history
- Polymerase chain reaction (PCR) testing for viral RNA, 4 days before their intended procedure
  1. The ratio between asymptomatic and symptomatic patients would be relatively constant, for a uniform demographic distribution
  2. Changes in PCR positivity among asymptomatic individuals would precede changes in symptomatic PCR-detected infections with a temporal delay
  3. Trends in asymptomatic SARS-COV-2 infections would predict the behavior of the virus within the community as a whole

Statistical adjustment

- Account for specificity and sensitivity of PCR tests (Gelman and Carpenter, 2020)
- Use patients' residence zip codes to define the target population living in the catchment area and link to geospatial summaries that are predictive of the infection risk
  1. Poststratify to the American Community Survey (ACS) approximating the target population based on: sex, age (0-17, 18-34, 35-64, 65-74, and 75+), race (white, black, and other), and zip code indicators
  2. The zip code level covariates include the percentage of individuals with college degrees, employment rate, median household income, average poverty level, average Area Deprivation Index (ADI) and percentage of urban tracts

Bayesian modeling of SARS-COV-2 incidence

- Let p_i be the probability that the individual i tests positive $y_i = 1$ and be a function of the unknown sensitivity δ, unknown specificity γ, and the true viral incidence π_i

$$y_i \sim \text{Bernoulli}(p_i),\ p_i = (1-\gamma)(1-\pi_i) + \delta\pi_i$$

- Incidence model

$$\pi_i = \text{logit}^{-1}(\beta_1 + \beta_2 * male_i + \alpha^{\text{age}}_{age[i]} + \alpha^{\text{race}}_{race[i]} + \alpha^{\text{time}}_{time[i]} + \alpha^{\text{zip}}_{zip[i]})$$

- Zip code level

$$\alpha^{\text{zip}}_z = \alpha^{\text{zip}} + \beta_2 \cdot urban_z + \beta_3 \cdot edu_z + \beta_4 \cdot pov_z + \beta_5 \cdot employ_z + \beta_6 \cdot inc_z + \beta_7 \cdot adi_z$$

Hierarchical prior specification

- Sensitivity and specificity

$$y_\delta \sim \text{Binomial}(n_\delta, \delta)$$

$$y_\gamma \sim \text{Binomial}(n_\gamma, \gamma).$$

We use the findings based on the Stanford Santa Clara study: $y_\delta/n_\delta \approx 0.7$ and $y_\gamma/n_\gamma \approx 1$

- Varying intercepts

$$\alpha^{name} \sim \text{normal}(0, \sigma^{name}),\ \sigma^{name} \sim \text{normal}_+(0, 2.5),$$

for name $\in age, race, zip$

$$\alpha^{time} \sim \text{normal}(0, \sigma^{time}), \sigma^{time} \sim \text{normal}_+(0, 5)$$

to allow a large variation

- Structured prior for high-order interaction terms (Si et al., 2020)

Poststratification

- Illustrating with the hospital test data, for each cell in the cross-tabulation table of sex (2 levels), age (5 levels), race (3 levels) and zip codes, we have the cell-wise incidence estimate $\hat{\pi}_j^t$, and population count N_j, where j is the cell index
- Calculate the weekly prevalence estimate in the population

$$\pi_{avg}^t = \frac{\sum_j N_j \hat{\pi}_j^t}{\sum_j N_j}.$$

- Calculate the weekly prevalence estimate across counties ($c = 1, \ldots, 98$)

$$\pi_c^t = \frac{\sum_{j \in c} N_j \hat{\pi}_j^t}{\sum_{j \in c} N_j}.$$

Example 3: Integrating RCT with a probability sample

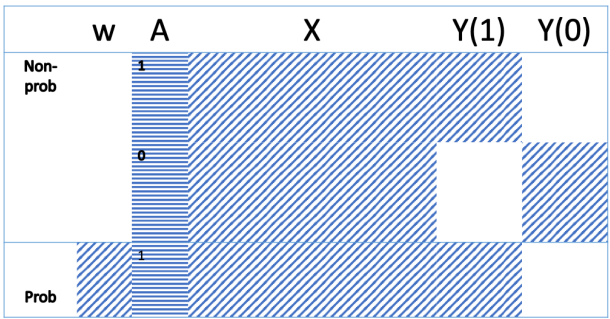


Figure: The probability sample includes cases under treatment.

Estimating the PATT

- Only using RCT: predicting potential outcome
- Using both RCT and the reference probability sample
  1. Combine IPW with prediction: two versions of quasi-weights

$$\Delta_{PATT} = \frac{\sum_{i=1}^{N} w_i^{np} I(S_i^{np} = 1) I(A_i = 1)(y_i - \hat{Y}_i(0))}{\sum_{i=1}^{N} w_i^{np} I(S_i^{np} = 1) I(A_i = 1)}$$

  2. Direct imputation: impute potential outcomes of RCT and the reference sample

$$\Delta_{PATT}^{pred} = \frac{\sum_{i=1}^{N} [I(S_i^{np} = 1) + (w_i^p - n_{np}/n_p) I(S_i^p = 1)] I(A_i = 1)(y_i - \hat{Y}_i(0))}{\sum_{i=1}^{N} [I(S_i^{np} = 1) I(A_i = 1) + (w_i^p - n_{np}/n_p) I(S_i^p = 1) I(A_i = 1)]}$$

- Example code available here:
  `https://github.com/yajuansi-sophie/generalizability`

Conclusions and discussions

- Data landscaping has been substantially changed
- Need novel collection processes and methodological developments
- Need to ensure equity across various segments of the population
- Address privacy concern and protect confidentiality, who gets access and for what purpose

References I

AAPOR Task Force (2023, Feb). Data quality metrics for online samples: Considerations for study design & analysis. https://aapor.org/reports/data-quality-metrics-for-online-samples-considerations-for-study-design-analysis/.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–972.

Broniecki, P., L. Leemann, and R. Wüest (2022). Improved multilevel regression with poststratification through machine learning (autoMrP). *Journal of Politics 84*(1), 597–601.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association 115*(532), 2011–2021.

Chipman, H. A., E. I. George, and R. E. McCulloch (2020). Bart: Bayesian additive regression trees. *Annals of Applied Statistics 4*, 266–298.

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

Covello, L., A. Gelman, Y. Si, and S. Wang (2021). Routine hospital-based SARS-CoV-2 testing outperforms state-based data in predicting clinical burden. *Epidemiology 32*(6), 792–799.

Dahabreh, I. J., S. E. Robertson, J. A. Steingrimsson, E. A. Stuart, and M. A. Hernan (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine 39*, 1999–2014.

Degtiar, I. and S. Rose (2023). A review of generalizability and transportability. *Annual Review of Statistics and Its Application 10*, 7.1–7.24.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice 2*(6).

Elliott, M. R., O. Carroll, R. Grieve, and J. Carpenter (2023). Improving transportability of randomized controlled trial inference using robust prediction methods. *Statistical Methods in Medical Research 32*, 2365–2385.

Elliott, M. R. and R. Valliant (2017). Inference for nonprobability samples. *Statistical Science 32*(2), 249–264.

References II

Fuller, W. A. and L. F. Burmeister (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association 29*, 247–249.

Gelman, A. and B. Carpenter (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society Series C (Applied Statistics) 69*(5), 1269–1283.

Gelman, A. and T. C. Little (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology 23*, 127–135.

Ghitza, Y. and A. Gelman (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science 57*, 762–776.

Gravano, L., P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srivastava (2001). Approximate string joins in a database (almost) for free. *Proceedings of the 27th International Conference on Very Large Data Bases*, 491–500.

Hartley, H. O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association 19*, 203–206.

Hernandez, M. A. and S. J. Stolfo (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery 2*, 9–37.

Holt, D. and T. M. F. Smith (1979). Post stratification. *Journal of the Royal Statistical Society Series A 142*(1), 33–46.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kastellec, J. P., J. R. Lax, M. Malecki, and J. H. Phillips (2015). Polarizing the electoral connection: Partisan representation in supreme court confirmation politics. *The Journal of Politics 77*(3), 787–804.

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

Kim, J. K., S. Park, Y. Chen, and C. Wu (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(3), 941–963.

References III

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology 32*(2), 133–142.

Kuriwaki, S., S. Ansolabehere, A. Dagonel, and S. Yamauchi (2023). The geography of racially polarized voting: calibrating surveys at the district level. *American Political Science Review*, 1–18.

Li, K. and Y. Si (2024). Embedded multilevel regression and poststratification: Model-based inference with incomplete poststratifier information. *Statistics in Medicine 43*(2), 256–278.

Mercer, A., A. Lau, and C. Kennedy (2018). For weighting online opt-in samples, what matters most? https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/.

Nandram, B. and J. Rao (2021). A bayesian approach for integrating a small probability sample with a non-probability sample. In *Proceedings of the American Statistical Association (JSM2021-Virtual Conference)*, pp. 1568–1603.

Rafei, A., C. A. C. Flannagan, and M. R. Elliott (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*(1), 148–180.

Reilly, C., A. Gelman, and J. Katz (2001). Poststratification without population level information on the poststratifying variable, with application to political polling. *Journal of the American Statistical Association 96*, 1–11.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey methodology 33*(2), 99–119.

Si, Y. (2024). On the use of auxiliary variables in multilevel regression and poststratification. *Statistical Science Forthcoming*.

Si, Y., L. Covello, S. Wang, T. Covello, and A. Gelman (2022). Beyond vaccination rates: A synthetic random proxy metric of total SARS-CoV-2 immunity seroprevalence in the community. *Epidemiology 33*(4), 457–464.

Si, Y., N. S. Pillai, and A. Gelman (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis 10*(3), 605–625.

Si, Y., T. Tran, J. Gabry, M. Morris, and A. Gelman (2024). Multilevel regression and poststratification interface: Application to track community-level covid-19 viral transmission. *arXiv preprint arXiv:2405.05909*.

References IV

Si, Y., R. Trangucci, J. S. Gabry, and A. Gelman (2020). Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology 46*(2), 181–214.

Si, Y. and P. Zhou (2021). Bayes-raking: Bayesian finite population inference with known margins. *Journal of Survey Statistics and Methodology 9*(4), 833–855.

Skinner, C. J. and J. N. K. Rao (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association 91*, 349–356.

Smeeding, T. M., D. S. Johnson, and C. F. Citro (Eds.) (2024). *Creating an Integrated System of Data and Statistics on Household Income, Consumption, and Wealth: Time to Build*. National Academies Press.

Steorts, R. C., R. Hall, and S. E. Fienberg (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association 111*, 1660–1672.

Stuart, E. A., C. P. Bradshaw, and P. J. Leaf (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. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology 8*(2), 231–263.

Wiśniowski, A., J. W. Sakshaug, D. A. Perez Ruiz, and A. G. Blom (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology 8*(1), 120–147.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Surv. Methodol 48*, 283–311.