Predictive Modeling Using an Enhanced Address-Based Sampling Frame

*Rachel Harter

Joe McMichael

Joint Statistical Meetings, August 11, 2022

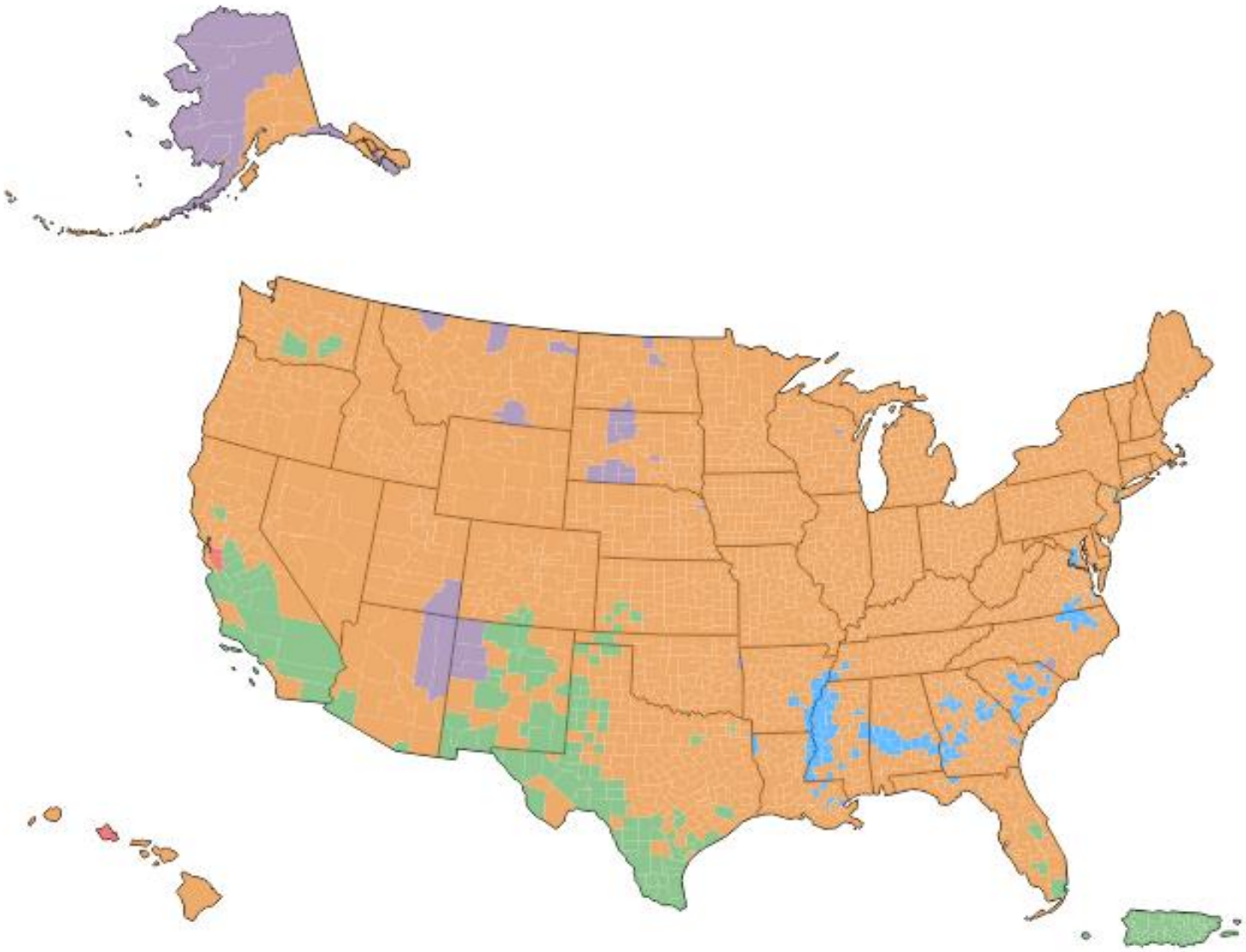
Washington, DC

RTI INTERNATIONAL

Sampling for a Subpopulation in ABS Designs

- Screen a general population sample of addresses
  - Expensive and inefficient
- Stratify and sample disproportionately
  - At least two strata: high-density and low-density
  - Sample at a higher rate in the high-density stratum
- ABS frames require auxiliary data for demographic stratification
- Geographic stratification is common
  - Geocode addresses and assign to geographical areas
  - Match to census or ACS data for the geographical areas

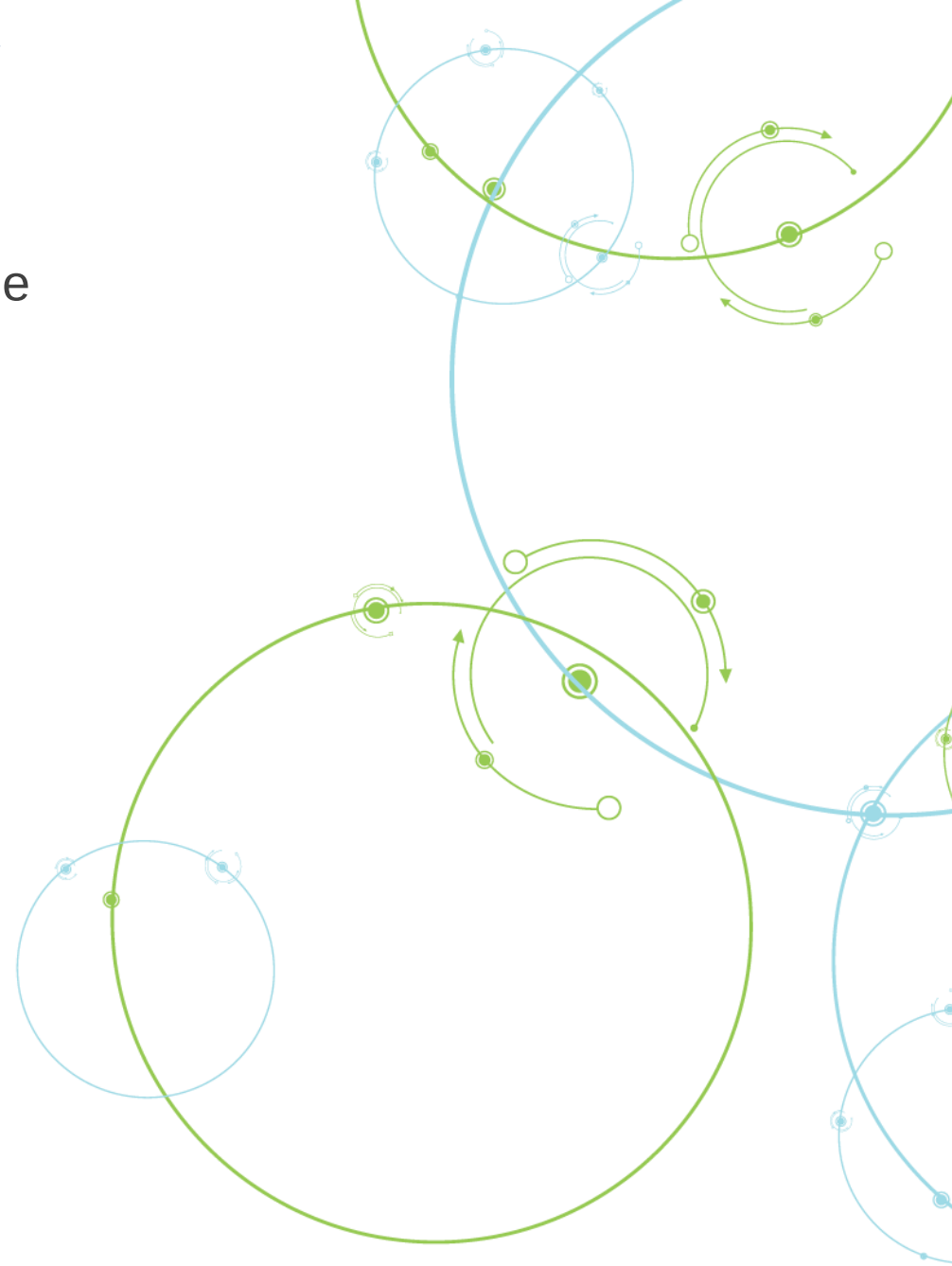
Example: Map of Density Strata by County



Options, Depending on Mode of Data Collection

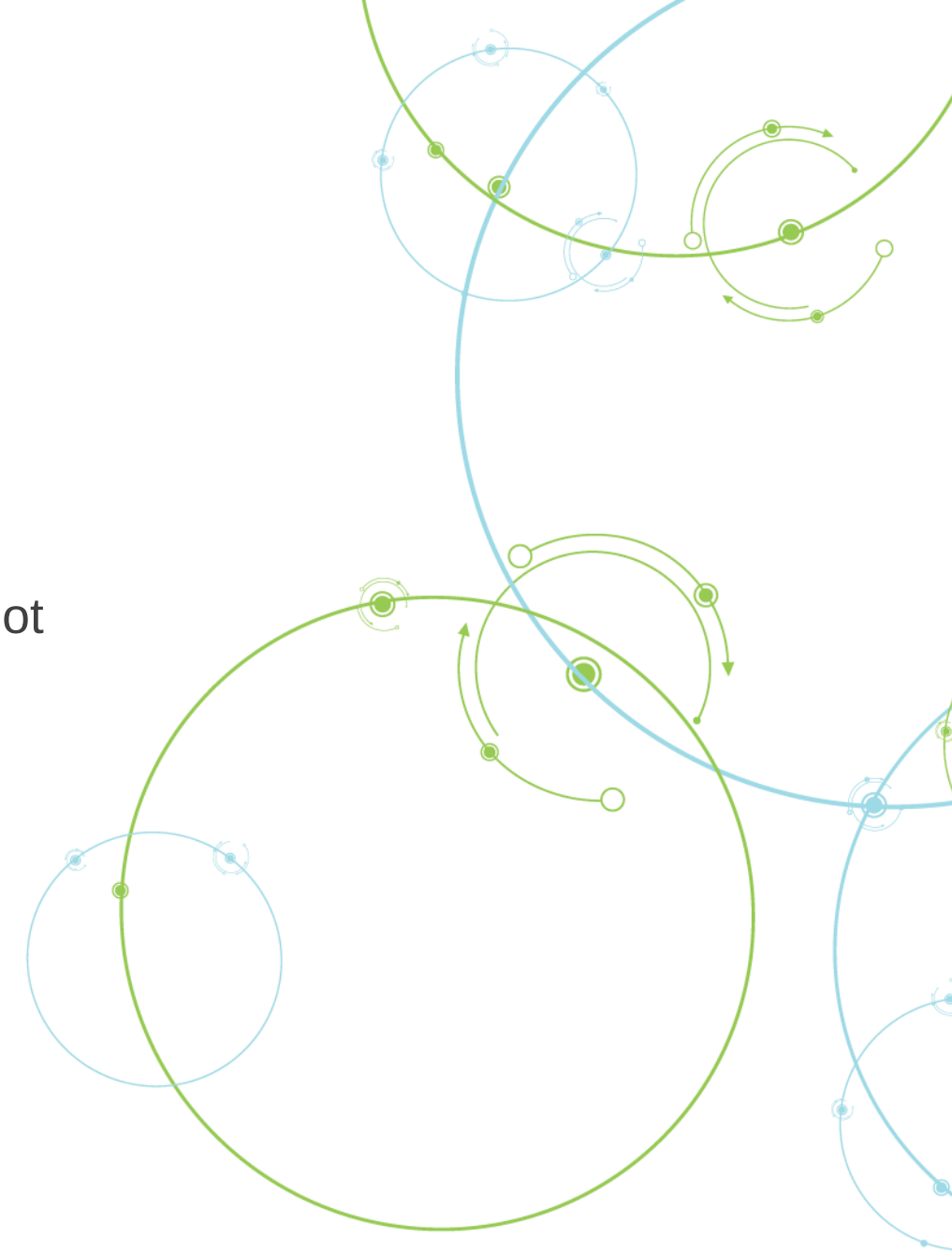
- Select geographic clusters from the strata first, then addresses within clusters.

- Select addresses directly from the strata.



Drawbacks

- Still might require more screening than desirable.
- Geographic stratification is not helpful if target subpopulation is not geographically clustered.



What if you could stratify at the address level?

Consumer Marketing Data Can be Useful

- Designed to reach as many potential customers as possible
  - (not survey research)

- Abundant
- Many demographics and behaviors available
- Often at the person level
- Often in the form of flags
- Generally expensive
- Often incomplete
- Often inaccurate

- Can roll-up to address level and match to ABS frame

- Match rate and accuracy can vary considerably (Harter 2016)

National Survey of Family Growth (NSFG)

(West, Wagner, Hubbard, & Gu 2015)

- Modeled eligibility for NSFG
- Covariates included
  - Frame variables
  - Marketing data from Aristotle
  - Marketing data from MSG's 3 unnamed vendors
- Models that included marketing covariates fit eligibility status much better than models using frame covariates alone.
- Models were applied in subsequent cycles of NSFG.

Predictive Modeling for Subpopulation Eligibility

(McPhee 2022)

- Fit a model of eligibility using training data.
- Test the model on separate data.
  - Use model predictions on the sampling frame to classify units as likely eligible or not.
  - Compare predictions to eligibility outcomes.

- The goal is prediction of subpopulation eligibility.
- Parameter estimates do not matter.

- Even a mediocre model may be useful.

RTI's Enhanced Frame

- Leased copy of the USPS Computerized Delivery Sequence file.
- Geocoded and assigned to census geographical areas
  - Appended area auxiliary variables from decennial census, ACS, and other federal sources
- Leased consumer marketing data.
  - Aggregated the person-level data to address level.
  - Merged address-level variables with the address frame.
- Having an in-house enhanced frame has supported extensive research. (https://abs.rti.org)

Review of RTI Sample Designs with Predictive Modeling

1. **New York Adult Tobacco Survey** for the NY State Dept of Health
   - Used data from prior cycle to predict prevalence of adult smokers for all census block groups
   - Stratified census block groups and oversampled in high-density strata

2. **Evaluation of Public Education Campaign on Teen Tobacco (ExPECTT)** to evaluate the FDA's youth tobacco prevention campaign.
   - Sample of 45,000 addresses stratified directly on an age group flag (not predictive modeling)

3. **Rural Smokeless Tobacco Education Campaign (RuSTEC)** evaluated another FDA campaign to prevent and reduce smokeless tobacco use among rural male youths 11-16.
   - Eligibility predicted from models developed with ExPECTT screener data.
   - RuSTEC data were used to develop models for two other studies.

Success of Predictive Modeling in RuSTEC

Ridenhour and McMichael (2017)

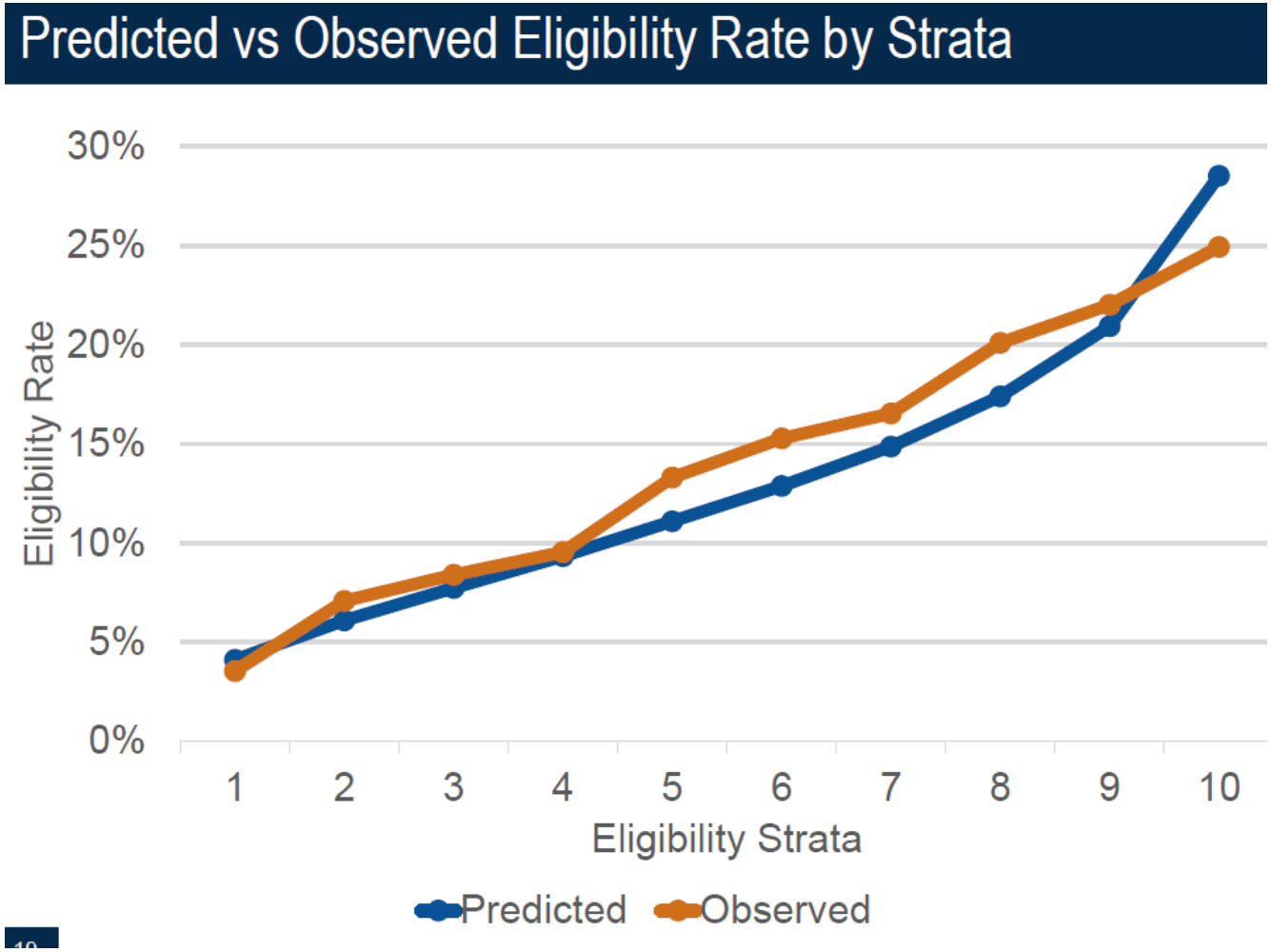
Stratum	Expected Eligibility Rate	Observed Eligibility Rate
1	2.9%	1.2%
2	4.4%	2.5%
3	9.3%	7.5%
4	13.3%	16.1%
5	25.6%	35.8%
6	30.0%	42.9%

Review of RTI Sample Designs with Predictive Modeling (cont.)

4. **ExPECTT-2**, which again targeted youths aged 11-16.

5. **Point of Sale Intervention for Tobacco Evaluation (POSITEv)** evaluated an FDA public education campaign around tobacco retail outlets.
   - Target population was households with a smoker aged 25-55.

POSITEv (McMichael & Wiant 2019)



Review of RTI Sample Designs with Predictive Modeling (cont.)

6. **National Recreational Boating Safety Survey** was a 2-part survey of recreational boating for the U.S. Coast Guard.
   - Two frames: Incomplete registry of boat owners and ABS
   - Auxiliary geospatial data - number of boats per geography
   - Not permitted to link geospatial data to frames
   - Model 1 on geospatial data to predict boats per census block group
   - Predictions used as auxiliary data for Model 2
   - Model 2 on registry data to predict boat ownership at address level
   - Model 2 applied to ABS frame for predictions of boat ownership

National Recreational Boating Safety Survey (Ridenhour et al. 2021)

Data Collection Rates* (%)

	Registry Frame	Stratified ABS Frame	Total
Screener	33.6	15.1	22.2
Eligibility	91.9	43.2	71.4
Yield	30.9	6.5	15.9

*Based on 9 of 12 completed cohorts

Review of RTI Sample Designs with Predictive Modeling (cont.)

7. **Recreational Boat Fishing Survey** - survey of people who fish by boat.
   - Subset of NOAA's Fishing Effort Survey to monitor recreational saltwater fishing activity by residents of Atlantic and Gulf Coast states.
   - ABS frame with state databases of licenses saltwater anglers.
   - Really rare subpopulation!

8. A current study to identify multigenerational households.

9. **National Survey of Family Growth** (again!) to find age groups

Risks of Stratification and Oversampling

- Oversampling for a subpopulation reduces the design's efficiency for total population estimates.
- Not sampling from the low-density stratum can lead to coverage bias.
- Selected target members in the low-density stratum will have larger weights, increasing DEFF and reducing effective n. It is possible to oversample the high-density stratum too much.
- Kalton (1986) gave formulas for optimal relative sampling rates

Guidelines for Successful Stratification and Oversampling

(Kalton 1986)

- Stratification and oversampling is more beneficial for more rare subpopulations.
- The high-density stratum needs to have *high* density of the targeted subpopulation.
- The high-density stratum needs to have a large proportion of the targeted subpopulation.
- The relative costs of the strata matter.

Additional Guidelines for Predictive Modeling for Stratification and Oversampling

- Predictive modeling needs good training data.
  - Repeated cross-sectional studies would be ideal.
- The auxiliary data (covariates) should be reasonably complete and accurate.
- The match rate of the auxiliary data to the frame should be high.
- The densities in the strata need to be known or estimated well.
- Even if not ideal, predictive modeling may be worth it.

Can You Stratify Without a Frame and Marketing Data?

- Yes, if you have a sample vendor that can match marketing data to sampled addresses.
  - Generally, vendors are prohibited from matching marketing data to the entire frame.

1. Obtain a large phase 1 sample.
2. Have the vendor match marketing data to the phase 1 sample.
3. If you have a model, apply it to the phase 1 sample. Otherwise, use marketing data directly.
4. Stratify the phase 1 sample.
5. Sample for phase 2 in the desired proportions.

References

Harter, Rachel (2016). The Quality of Auxiliary Variables in an Enhanced Address-Based Sampling Frame. Invited presentation in JSM Proceedings, Government Statistics Section, pp. 74-89, Alexandria: American Statistical Association.

Kalton, G. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society, Series A, Vol. 149, part 1, pp. 65-82.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Levine, B. (2016). "The Overpromise of Oversampling," Poster presented at AAPOR, Austin, TX.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

McMichael, J. & Wiant, K. (2019). Improvements in sample design with address-level prediction models. Presented at AAPOR, Toronto.

References (cont.)

McPhee, C. (2022). Applications of Predictive Modeling to Survey Design & Operation in Address-based Samples. Webinar presented March 17, 2022. American Association of Public Opinion Research.

Ridenhour, J. L. & McMichael, J. P. (2017). *Propensity stratification with auxiliary data for address-based sampling frames.* Presented at the American Association for Public Opinion Research conference, New Orleans, LA.

Ridenhour, J., McMichael, J., Harter, R., & Dever, J. (2014). ABS and Demographic Flags: Examining the Implications for Using Auxiliary Frame Information." Presented at the Joint Statistical Meetings, Boston, August 7, 2014.

Ridenhour, J., McMichael, J., Krotki, K., & Speizer, H. (2021). Using big data to improve sample efficiency. In *Big data meets survey science* (C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov & L. E. Lyberg, eds.). https://doi.org/10.1002/9781118976357.ch17

References (cont.)

West, B.T., Wagner, J., Hubbard, F., and Gu, Haoyu (2015). The Utility of alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, 3, 240–264.

Thank you

Contact: Rachel Harter – rharter@rti.org

Joseph McMichael – mcmichael@rti.org