Leveraging Survey Methods to Improve Administrative Record Estimates

Benjamin Reist

2nd FCSM/WSS Workshop on Quality of Integrated Data
01/25/2018

Can surveys be used to improve administrative record estimates?

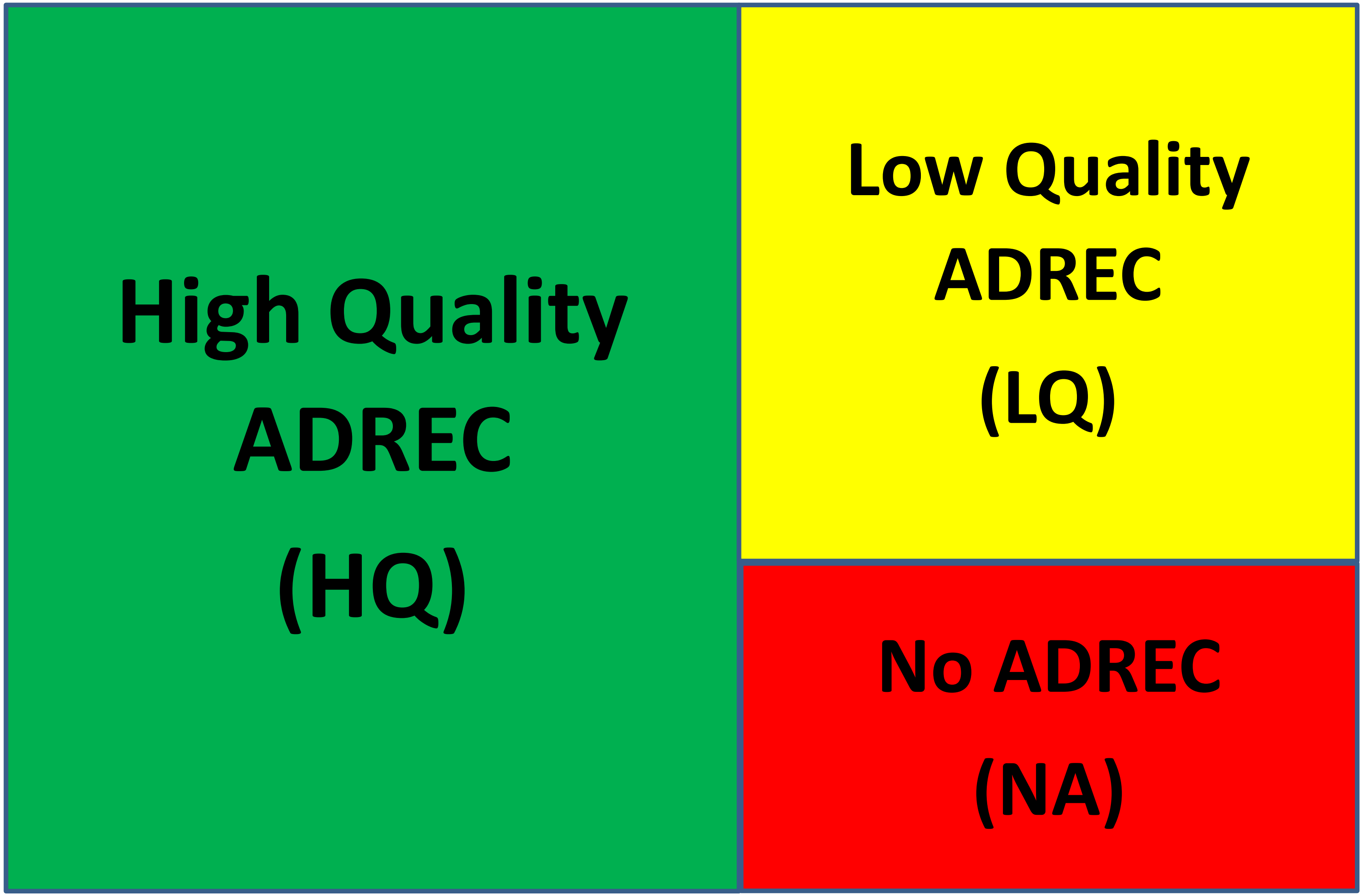
(current approach: use ADRECs to improve survey estimates)

Using ADRECs to Improve Survey Estimates

- Sample Design
- Data Collection Monitoring
- Estimation
  - Weighting
  - Editing/Imputation
  - Substitution
- Evaluations
  - Nonresponse bias studies
  - Measurement error evaluations

Possible Data Quality Issues Associated with ADREC Estimates

- Coverage
  - Records on sampling frame not in administrative records
- Measurement
  - Unknown measurement issue
  - Difference in what is measured
    - Time lag
    - Similar but different definitions



High Quality ADREC (HQ)

Low Quality ADREC (LQ)

No ADREC (NA)

How a Survey Could Help

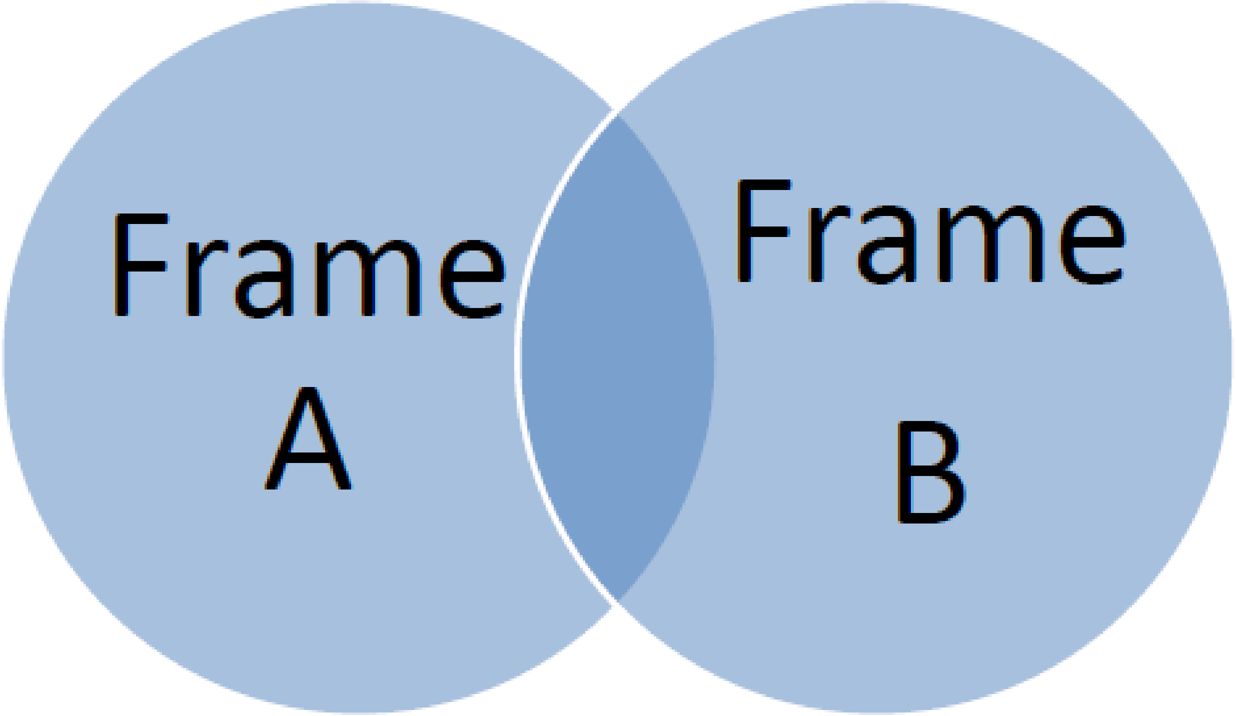
- Coverage
  - Estimate contribution of cases with no ADRECs

- Measurement
  - Adjust ADREC estimates to address measurement error
  - Monitor for new measurement issues in ADRECs

Survey Methods That Could be Used

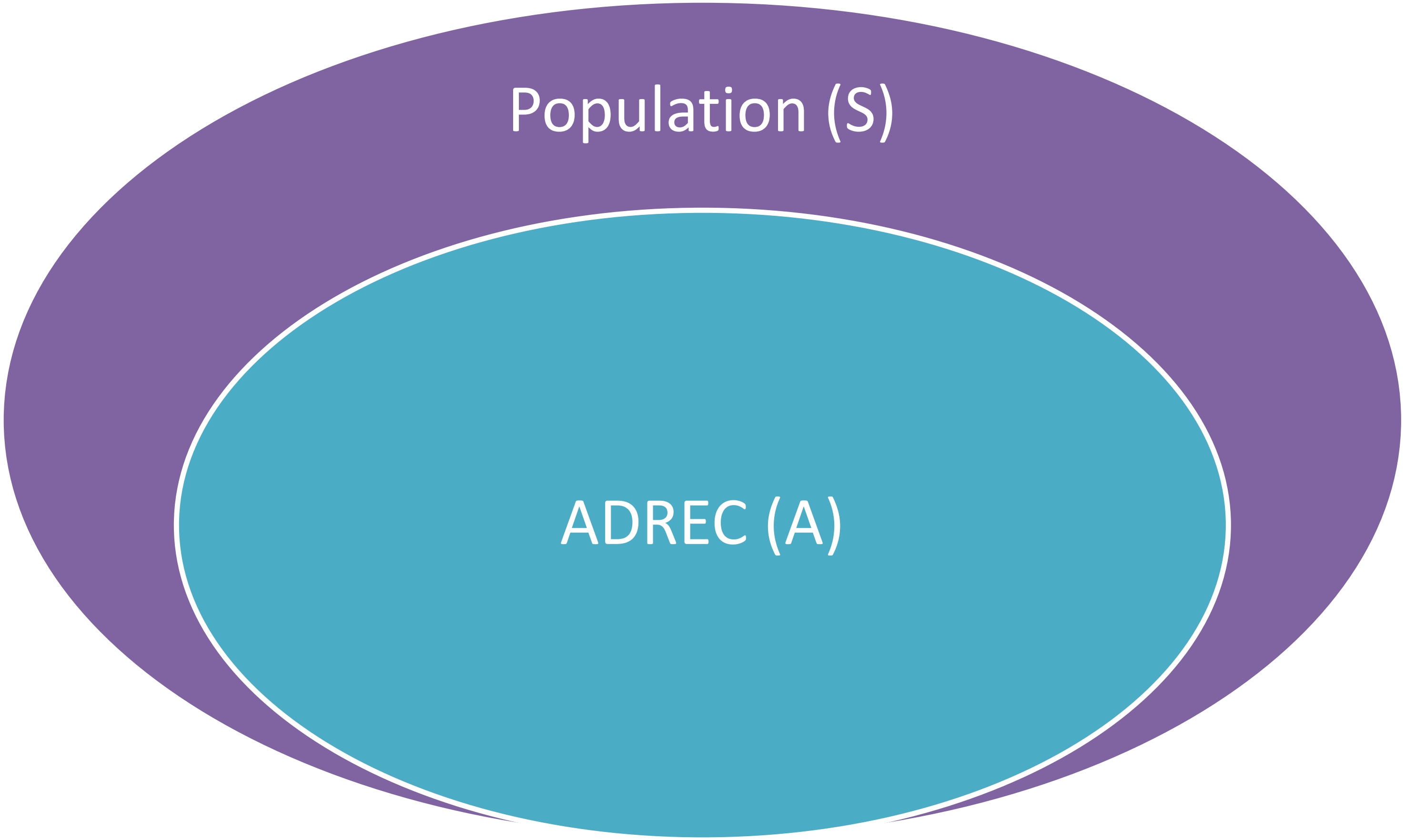
- Overlapping frames methodology
- Model-assisted estimation
  - Generalized Difference Estimators

Overlapping Frame Method



$$\hat{t} = \hat{t}_a + \lambda\hat{t}_{ab}^{A} + (1-\lambda)\hat{t}_{ab}^{b} + \hat{t}_b$$

Overlapping Frames for ADRECs



$$\hat{t} = \lambda \hat{t}_a^A + (1-\lambda)\hat{t}_a^S + \hat{t}_s$$

Measurement Error Model

- Additive error model assumed for ADREC estimate

$$\hat{t} = \lambda(\hat{t}_a^A + \delta_a) + (1 - \lambda)\hat{t}_a^S + \hat{t}_s$$

- δ_a is the bias in the ADREC estimate

Bias Estimation

- If the survey is assumed to be the "gold standard," using direct substitution

$$\hat{\delta}_a = \sum_{i=1}^{n} w_i (y_i^S - y_i^A)$$

- $\hat{\delta}_a$ is the survey estimate of the error in the ADREC estimate

Adjusted ADREC Estimate

- Combining the coverage and measurement error adjustments

$$\hat{t} = \lambda\left(\hat{t}_a^A + \sum_{i=1}^{n} w_i\left(y_i^S - y_i^A\right)\right) + (1-\lambda)\hat{t}_a^S + \hat{t}_s$$

- The first term can be thought of as a GREG estimator with intercept 0 and slope 1

Gold Standard Assumption

- Survey as the gold standard
  - Strong assumption
  - Wrong in many cases

- If ADREC is assumed to be the goal standard, then

$$\hat{t} = \hat{t}_a^A + \hat{t}_s$$

Assuming No Gold Standard

$$\hat{t} = \omega_s(\lambda \hat{t}_a^{GREG} + (1 - \lambda)\hat{t}_a^S) + \omega_A \hat{t}_a^A + \hat{t}_s$$

- ω_s is the probability the survey is correct
- ω_A is the probability the ADREC is correct
- $\omega_s + \omega_A = 1$
- $\hat{t}_a^{GREG} = \hat{t}_a^A + \sum_{i=1}^n w_i(y_i^S - y_i^A)$

Further Refinements

- Varying the λ and ω by domain
- Models other than direct substitution
  - Generalized Difference Estimators
    - GLM
    - Nonparametric Models
    - Time-to-Event Models
- Extends to multiple ADREC sources

Open Questions

- How to deal with nonresponse?
- How to allocate sample optimally across the domains and the part of the frame that is not covered by the ADRECs?
    - adaptively by rolling out sample in waves?
- How can this be done in a multivariate setting where there are multiple estimates of interest?

New Role For Data Collection

- To assist in estimating and updating the probability that the administrative records are correct in each domain
- To adjust bias caused by under-coverage and measurement error in administrative record estimates
- Monitor where administrative records could be improved

Contact Information

Benjamin.M.Reist@Census.gov