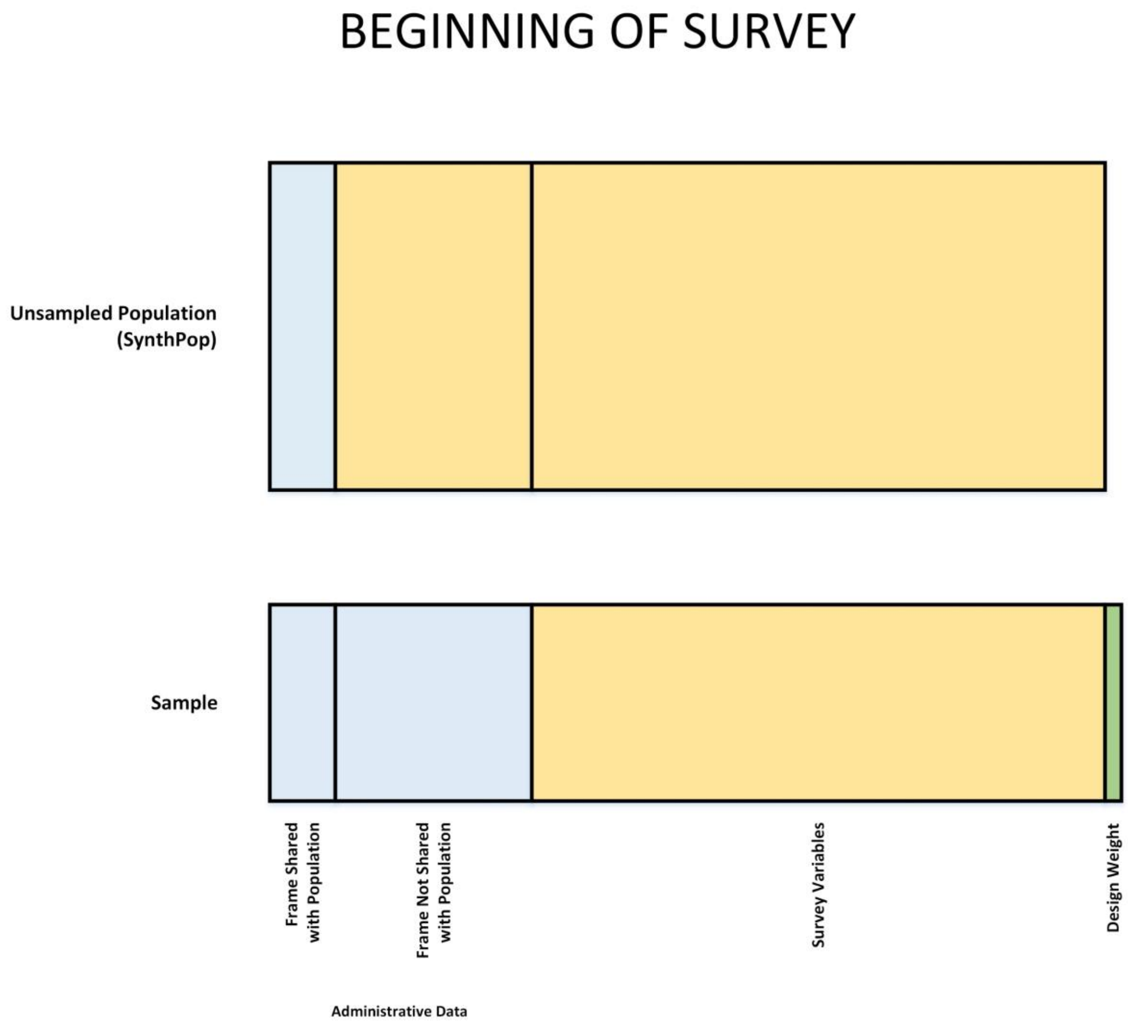
Predictive Modeling as an Alternative to (Re-) Weighting

ITSEW 2019, Bergamo, June 12, 2019

Alan F. Karr
RTI International

Surveys as Prediction Problems: Before



Introductory Example

- Pharmaceutical industry client wanted *full national dataset* with
  - Demographics, available from ACS = American Community Survey *for a sample of people:* ~15M in 5-year compilation
  - 23 variables relating to T2DM = Type II diabetes mellitus, available from NHANES = National Health and Nutrition Examination Survey *for a sample of people:* ~10,000 each year

- Why? Calculate Gini indices of representativity in clinical trials, once multiple inclusion and exclusion criteria are imposed

- Problem: No versions of Gini indices are available for weighted data

Simple (-Minded) Strategy 1

Cloning: Using *fully imputed* and mildly filtered NHANES dataset (n = 9813), create a dataset in which each record appears as many times as its weight

- Use fractional part of weight as probability to include one more copy
- Resultant dataset has 311,204,241 records

Simple (-Minded) Strategy 2

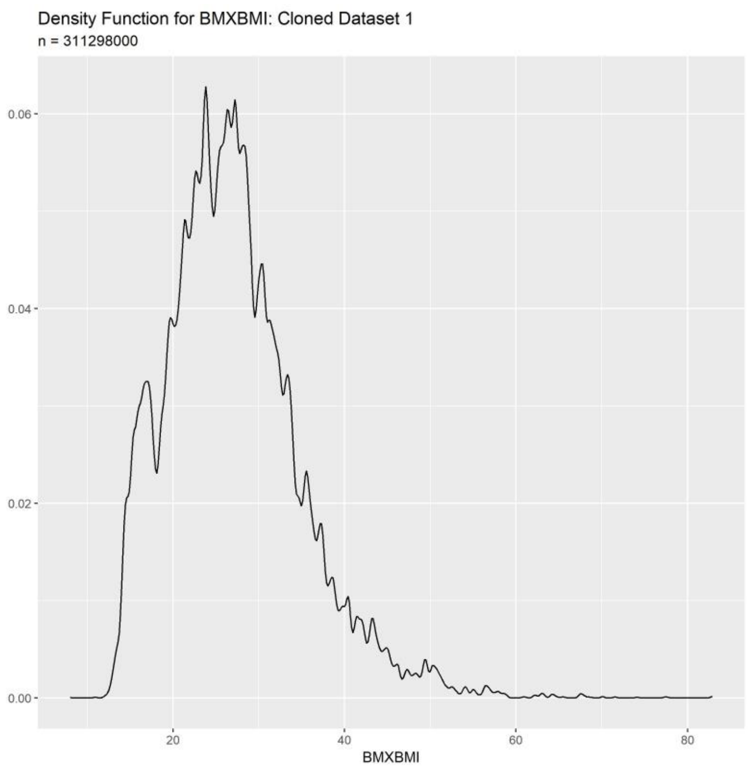
Resampling (Linkage)

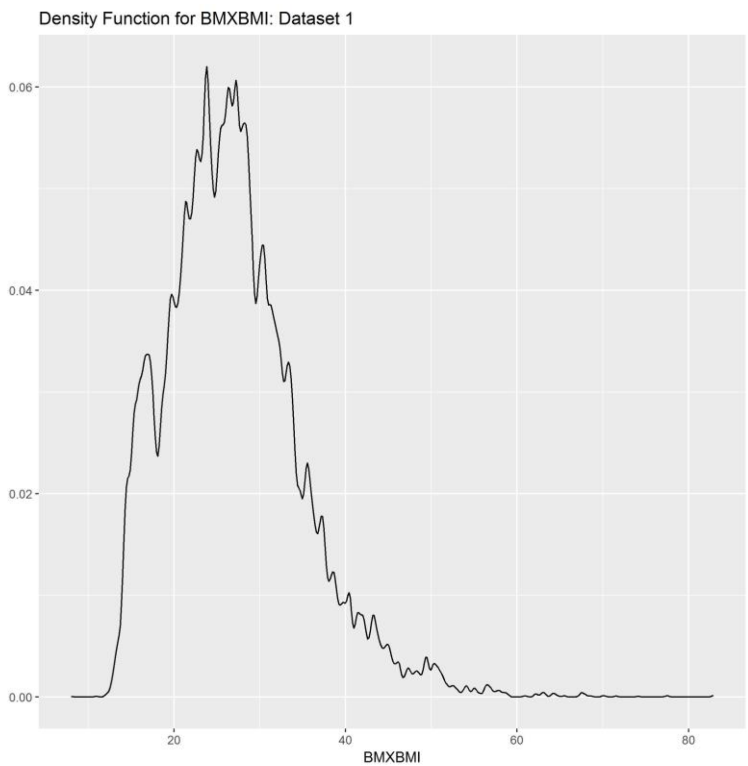
- **SynthPop:** Create a version of the RTI Synthetic Population containing 299,444,439 records and all ACS variables. Cross-tab of age and gender:

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70+
M	19,848,374	20,470.012	19.381,292	19,791,339	21,388,714	21,783,825	16,053,011	15,604,110
F	20,699,392	21,390,818	18,547,929	18,242,327	19,868,352	21,181,040	14,473,830	11,784,074

- **MADIS** (Model-Assisted Data Integration System) **Light:** For each cell in this table, sample that many records from the subset of the NHANES dataset that match on age and gender, using probabilities proportional to weights

Sample Results: No Meaningful Difference





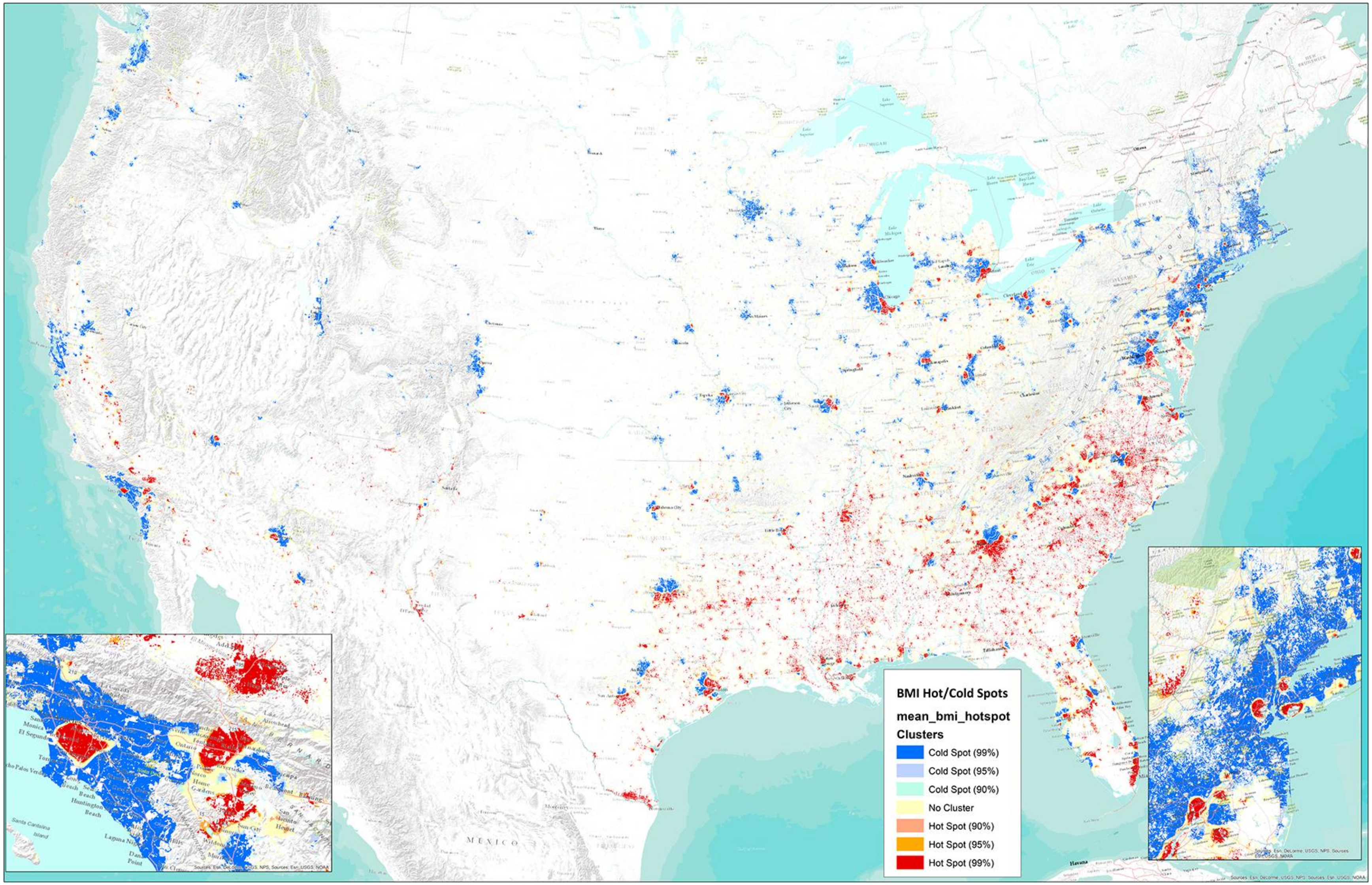
What's the Real Issue with these Strategies?

- Not enough variability! Every combination of the 23 NHANES variables in the synthesized national dataset appears intact in the NHANES dataset.

- Question to Ponder: Uncertainty quantification. Sources include:
  - Sampling and other forms of TSE in ACS
  - Sampling and other forms of TSE in NHANES, as well as added uncertainty from imputation
  - Cloning or resampling that creates national dataset

- This year's candidate for a new form of TSE: data augmentation error

A Step in the Right Direction: 2015 Obesity Data Challenge



Behind the Curtain

- **SynthPop** of ~ 200M adults containing 4 categorical predictors: age, gender, race/ethnicity, educational attainment, matched to released totals at block group level + block group geography

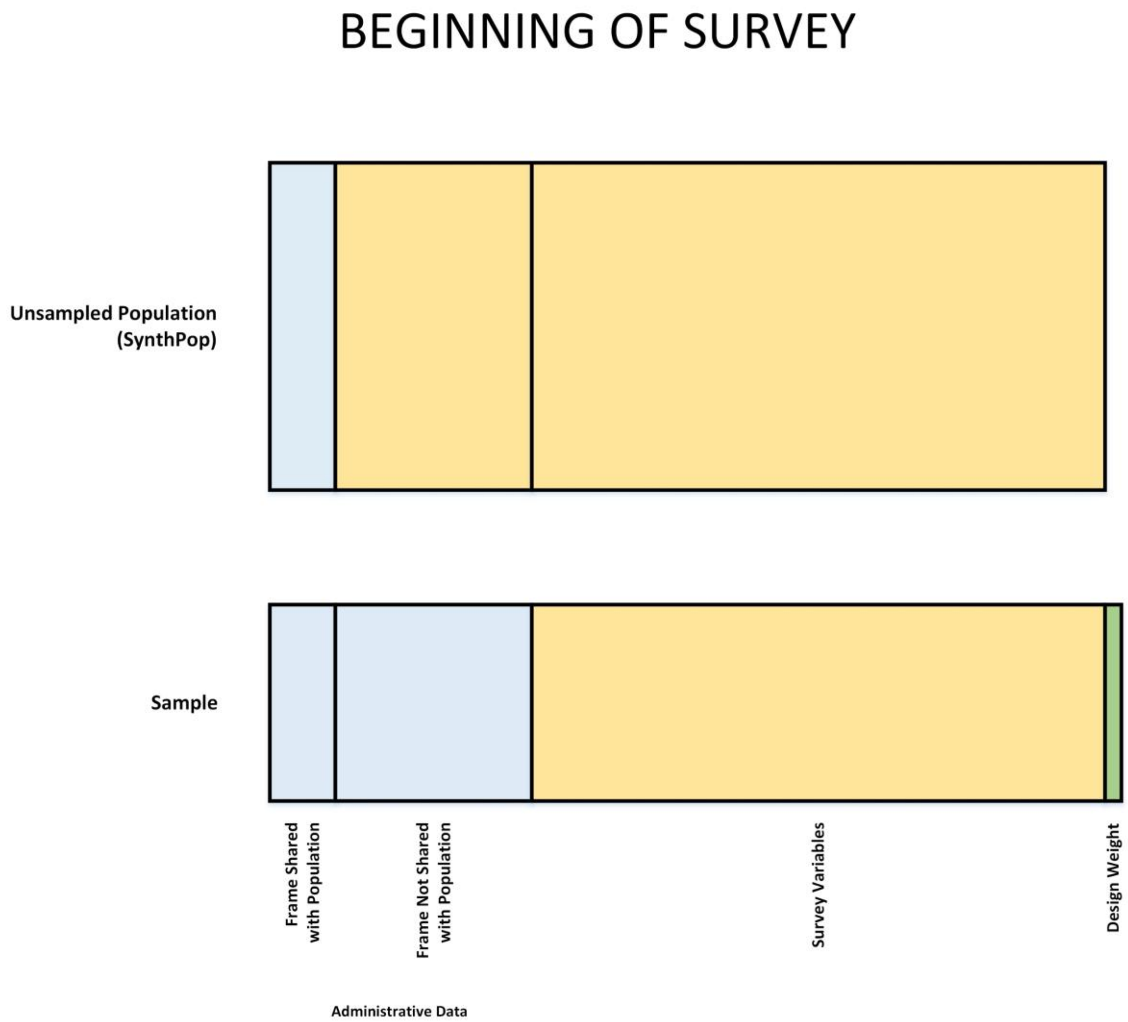
- **NHANES** dataset containing the 4 predictors and BMI

  [Data harmonization]

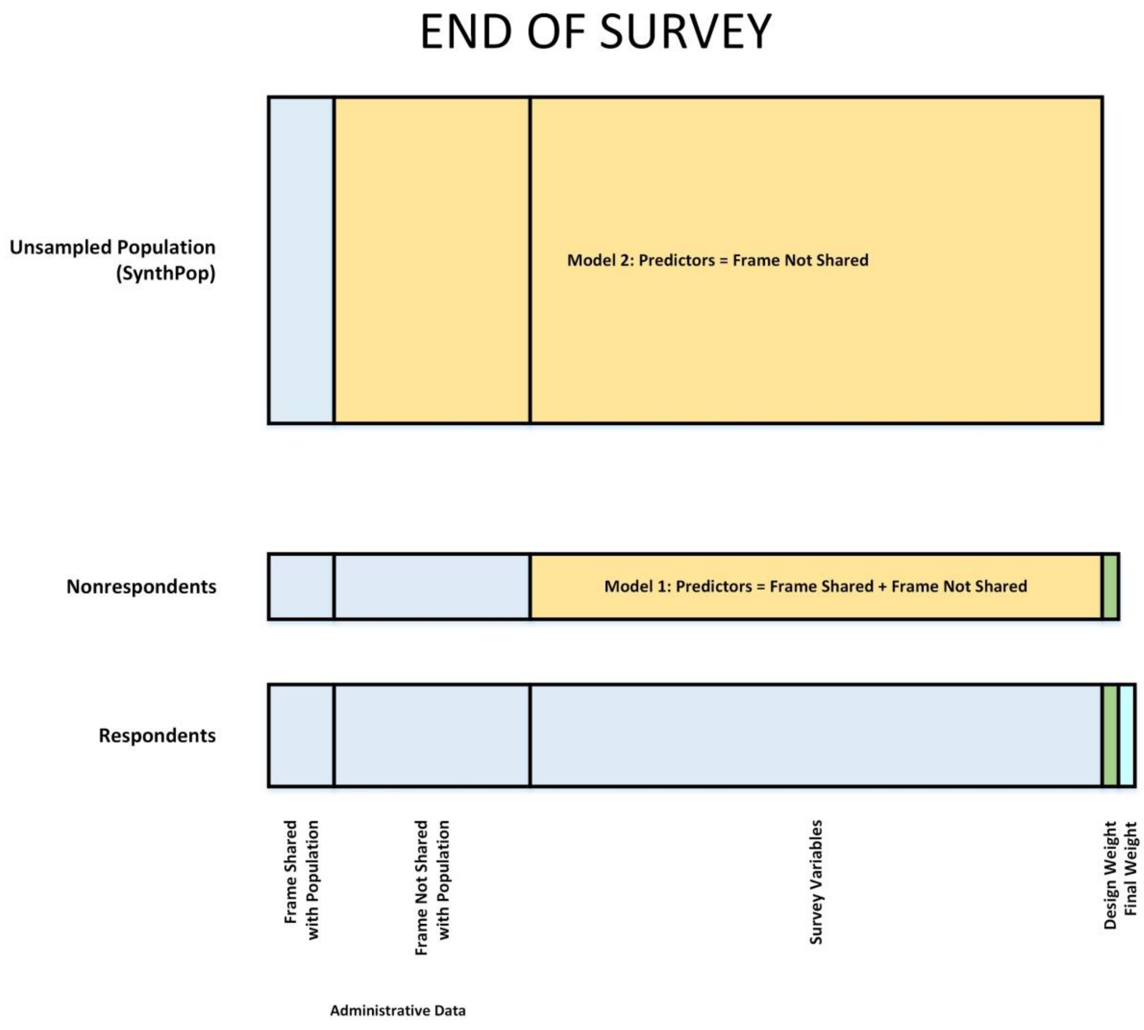
- **Log-normal models for BMI:** 95 distributions, one for each combination of the 4 predictors (one combination collapsed into another because of insufficient sample)

- **Simulation:** for each SynthPop record, simulate a value of BMI from the associated log-normal distribution
  - Produces values of BMI that do not appear in the NHANES data!

Surveys as Prediction Problems: Before



Surveys as Prediction Problems: After



An Example (avoids reweighting)

- 2013 Medical Expenditure Panel Survey, Household Component (n = 26,863)
- Simulate nonresponse using FamilyIncome and TotalExpenditure
  - 22,209 respondents, 4654 nonrespondents
- Shared frame: Age, Gender, Race/Ethnicity, Region [in US]
  - [Recoding]
- Unshared frame: EducationalAttainment, HealthInsurance, MaritalStatus, FamilyIncome
  - [Recoding]
- Partition modeling (subsequently, in other contexts, nonparametric density estimation) *with weights* used to reconstruct survey variables for nonrespondents
  - Presence of any of 5 diseases (arthritis, asthma, CHD, diabetes, high cholesterol)
  - BMI
  - TotalExpenditure (interesting because of atom at 0)

Sample of Results: Any of 5 Diseases

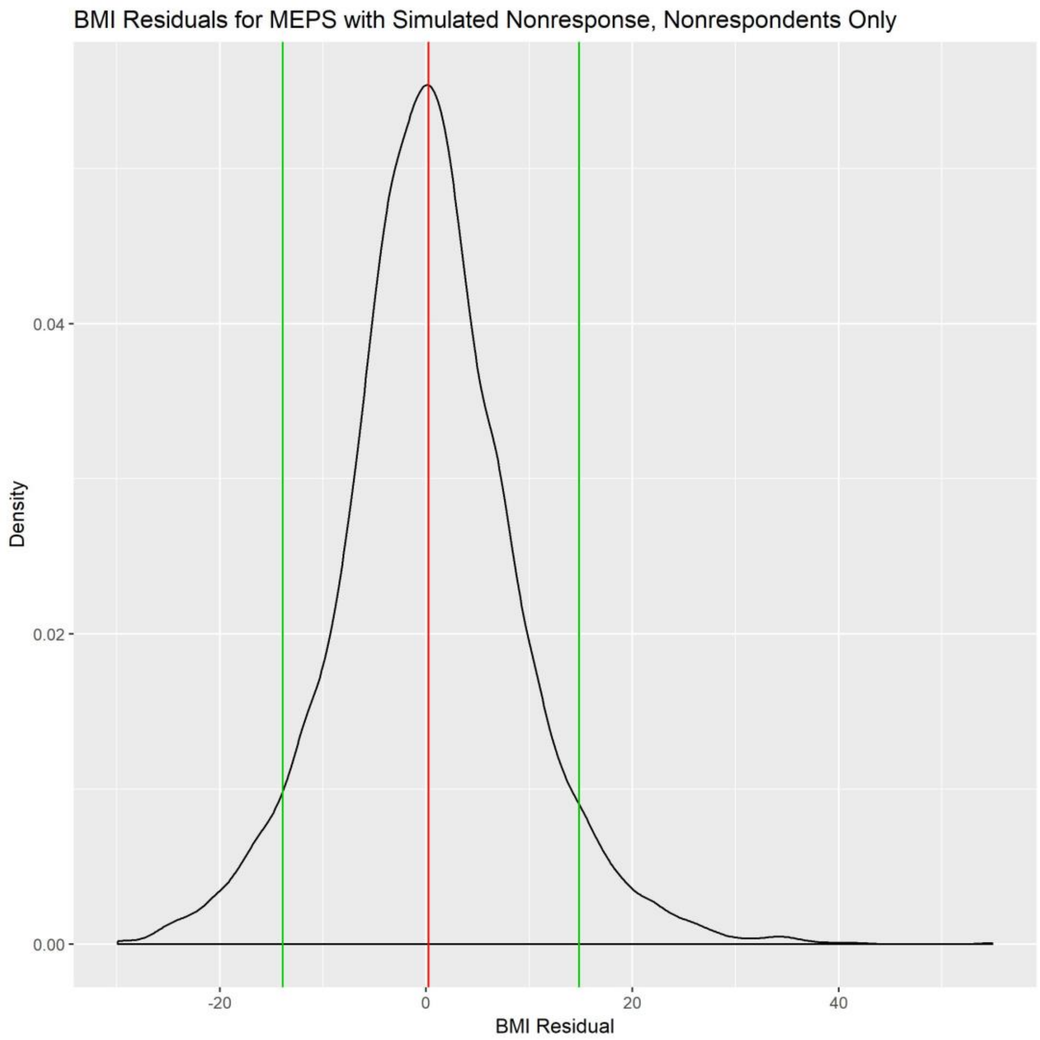
- Truth

Respondent?	Any Of Five	None	Sum
N	2976	1678	4654
Y	9261	12948	22209
Sum	12237	14626	26863

- Predictions

Respondent?	Any Of Five	None	Sum
N	2692	1962	4654
Y	9261	12948	22209
Sum	12237	14626	26863

Example of Results: BMI

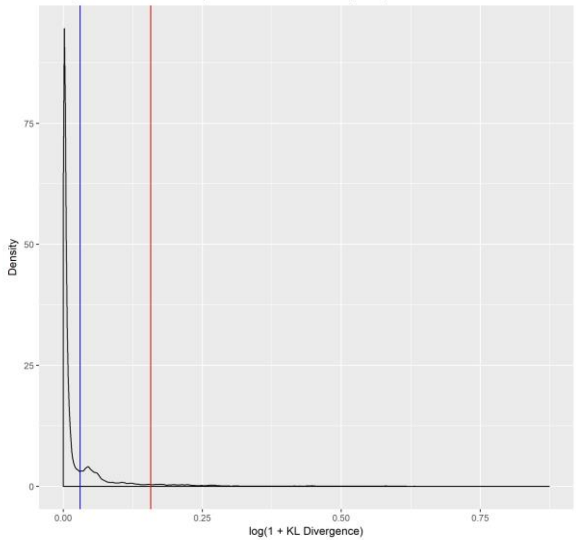


Where Things Stand

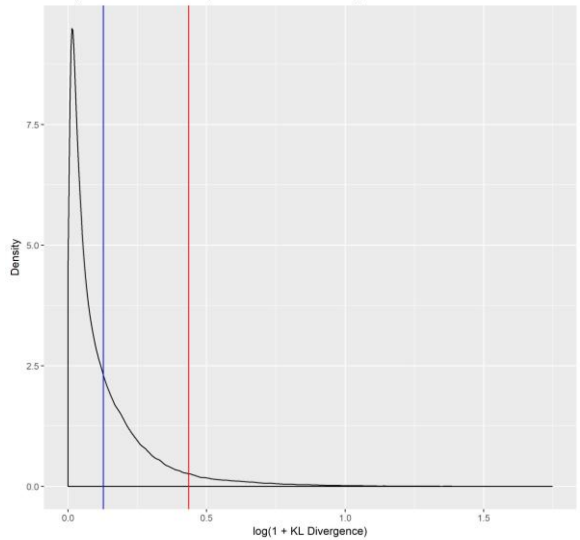
- Re-weighting amounts to cloning respondent records
- Resampling/linkage can be useful, but still cannot create records not present in the respondent data
- Modeling has the potential to
  - Create much richer datasets
  - Increase usability (initial example)
- It is possible to account for modeling-induced uncertainty
  - Observed to date: modeling variability is often comparable to sampling variability

Pondering UQ

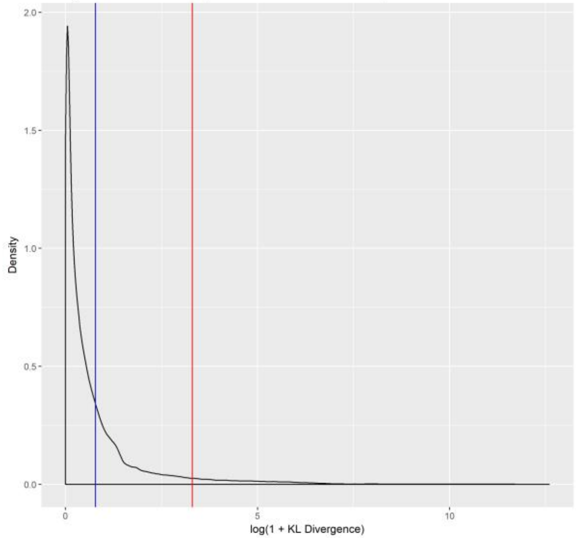
Kullback-Leibler Divergence of Joint Distribution of
Family Income and Total Expenditure over 369 Linkage Replications



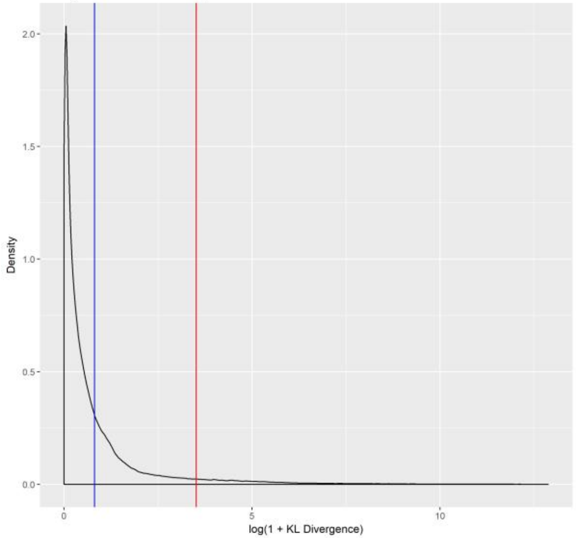
Kullback-Leibler Divergence of Joint Distribution of
Family Income and Total Expenditure over 500 Linkage Simulations



Kullback-Leibler Divergence of Joint Distribution of
Family Income and Total Expenditure over 369 MADIS Replications



Kullback-Leibler Divergence of Joint Distribution of
Family Income and Total Expenditure over 500 MADIS Simulations



Unresolved Challenges

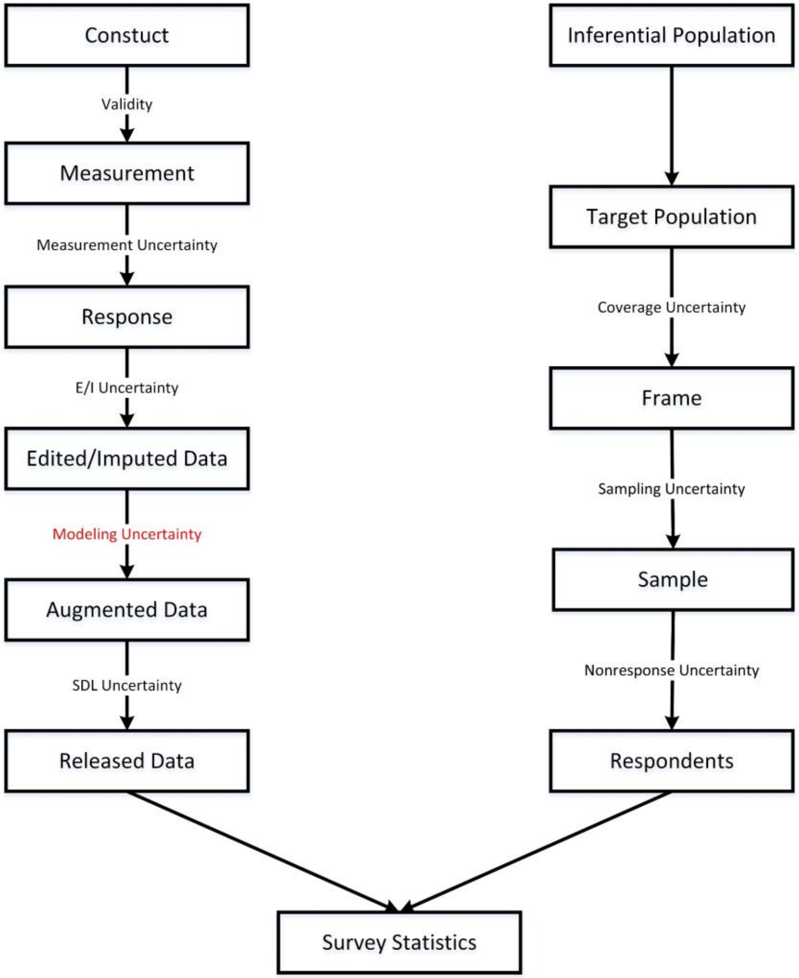
- Modeling lacks transparency and reproducibility
  - “Trust us, we’re smart”
  - Adding multiple variables requires conditional independence assumptions that are hard to verify

- Too much of the modeling process is manual, therefore not scalable
  - Identification of variables that match
    - May be resolvable via AI and high-quality metadata
  - Harmonization
  - Order of addition of variables

- Model validation
  - Simulation of additional nonresponse is a good potential strategy

- **Uncertainty quantification**

Parting Shot



Total Survey Uncertainty

RTI International

delivering **the promise of science**
for global good

RTI INTERNATIONAL

Alan F. Karr, Director, CoDA
karr@rti.org
+001 919 316 3423