Utility of Commercial Data for Sampling Population Subgroups

A Case of Health and Retirement Study

MPSDS-JPSM Brownbag Seminar
11/16/2022

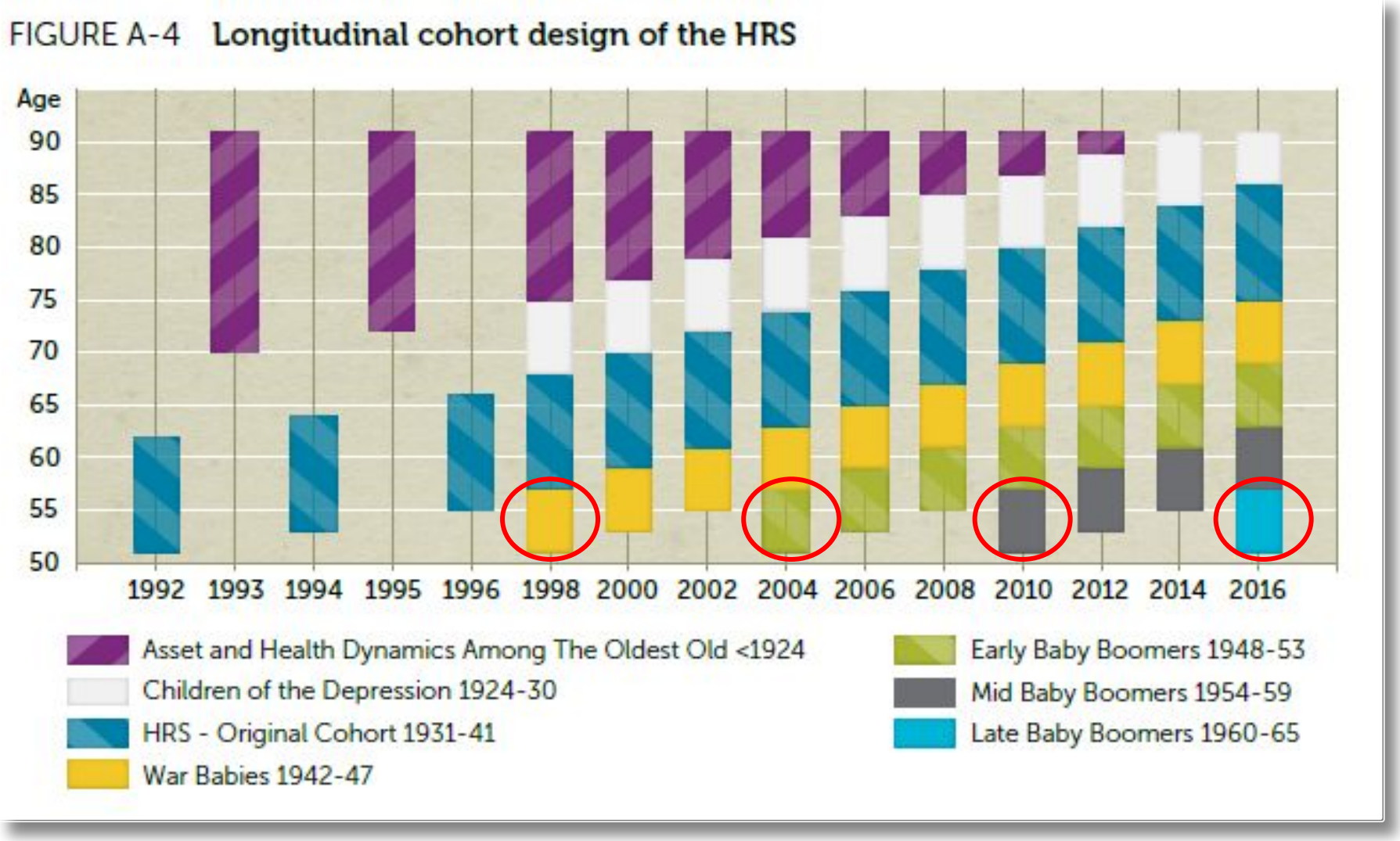
Sunghee Lee, Chendi Zhao, Anqi Liu

INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

Health and Retirement Study (HRS)

- Flagship longitudinal aging study in the U.S.
- Target the population aged ≥50 y.o.
- Focus on health and finance near and through the retirement stage
- Started in 1992 with its original age cohort born 1931-1941
- Other age cohorts added over time
- Since 1998, a new age cohort recruited every 6 years to represent ≥50 y.o.

HRS Cohort Design



Source: https://hrsonline.isr.umich.edu/sitedocs/databook/

HRS Sampling For New Cohort Recruitment

Traditionally,

- 3-stage stratified area probability sampling
  - Primary sampling units

    Metropolitan statistical area or counties; stratified by certainty; *pps* with the age eligible population size as a measure of size (MoS) within stratum
  - Secondary sampling units

    Groups of census blocks; stratified by race/ethnicity distribution; *pps* with the age eligible population size as MoS within stratum
  - Addresses

    Screen for age eligible financial units

HRS 2016 Sampling

- Recruitment of the late baby boomers (LBB) cohort born 1960-1965; oversample minority
- Stratification of addresses enhanced by commercial data on age and race/ethnicity
  - LBB - Black
  - LBB - Hispanic
  - LBB - Other race/ethnicity
  - Not LBB
  - No age information
  - No commercial data

  → Higher selection probabilities to LBB and racial/ethnic minority address strata

HRS 2016 Address Stratification Results

	n	Screener completed (%)	Age eligible among screener completed (%)	Black/Hispanic among age eligible (%)
Total	**54,066**	**60.2**	**15.5**	**40.6**
By address strata				
LBB - Black	4,396	66.3	31.0	69.2
LBB - Hispanic	2,818	63.1	32.4	79.2
LBB - Other	11,055	57.4	32.9	12.4
Not LBB	19,452	66.0	5.2	46.0
No age	4,129	57.7	9.5	36.7
No commercial data	12,216	51.5	9.4	54.0

What is this study about? – 1

1. Address frame analysis (n=170,435)
2. Screener sample analysis (n=54,066)
3. Screener respondent analysis (n=33,576)
4. Main survey financial unit analysis (n=3,189)

1~4: Commercial data **availability** ‡

~ Sample design + ACS data§ (+ additional predictors)

4: Commercial data **accuracy** ‡

~ Sample design + ACS data§ + Survey data

‡ Predictors selected through Bayesian additive regression trees

§ American Community Survey 2013-2017 5-yr Summary File at the census block-group level

What is this study about? – 2

- Commercial data availability
  - Any data, Income, Age, Race
  - Data from every quarter 2015-2017

    ⟶ First quarter of 2016 (i.e., not the data used to create the address strata)
  - Information from two different vendors combined

    ⟶ Available from either vendor

What is this study about? – 3

- Commercial data accuracy
  - Conditional upon commercial & survey data availability
  - Match between commercial data and screener/main survey data
  - Income: <$50K vs. ≥$50K
  - Race: Minority vs. Non-Minority

Address Frame Analysis

Results: Frame – Commercial Data Availability

(n=170,435)	Any	Income	Age	Race
% available	78.2	78.2	59.6	64.7
Predictors	**OR**	**OR**	**OR**	**OR**
PSU: Certainty vs. Non-certainty	1.19***	1.19***	1.11***	1.09***
SSU Strata (ref: Non-minority)				
High Black	0.93**	0.93***	0.96*	1.06***
High Hispanic	0.89***	0.89***	0.93***	0.99
High Black+Hispanic	0.89***	0.89***	0.91***	0.99
Census Region (ref: Northeast)				
Midwest	1.48***	1.47***	1.12***	1.28***
South	1.53***	1.53***	1.12***	1.44***
West	1.48***	1.48***	1.13***	1.23***
ACS: % Occupied housing units	1.03***	1.03***	1.02***	1.02***
ACS: % Renter occupied housing units	0.99***	0.99***	0.99***	1.00***
ACS: % Single person housing units	1.01***	1.01***	1.01***	1.01***

*Note: *p<0.05, **p<0.01, ***p<0.001*; Other predictors not shown in the table include ACS % person speaking other than English, % Mobile home housing units and self response rates.

Screener Sample Analysis

Results: Scrn Sample – Data Availability

(n=54,066)	Income	Age	Race
% available	80.5	64.5	67.9
Predictors	**OR**	**OR**	**OR**
PSU: Certainty vs. Non-certainty	1.28***	1.04	1.04
SSU Strata (ref: Non-minority)			
High Black	0.84**	0.95	1.15**
High Hispanic	0.77***	0.93*	0.90*
High Black + Hispanic	0.72***	0.90**	0.89*
Census Region (ref: Northeast)			
Midwest	1.46***	0.91	0.98***
South	1.35***	0.90**	1.20***
West	1.34***	0.95	1.00***
HRS Screener: Response vs. Nonresponse	1.66***	1.26***	1.36***

*Note: *p<0.05, **p<0.01, ***p<0.001*; Other predictors not shown in the table include address strata, ACS variables.

Screener Respondent Analysis

Results: Scrn Respondents – Data Availability

(n=33,576)	Income	Age	Race
% available	84.1	67.8	71.6
Predictors	**OR**	**OR**	**OR**
HRS Screener: Cohort (ref: LBB)			
Older than LBB	1.15	1.24***	0.86*
Younger than LBB	0.60***	0.52***	0.58***
No cohort information	0.82*	0.81*	0.72***
HRS Screener: Race (ref: Other)			
Hispanic	0.69***	0.69***	0.84*
Black	0.83*	0.85*	0.93
Missing (=not LBB)	0.54***	0.50***	0.66***

*Note: *p<0.05, **p<0.01, ***p<0.001*; Other predictors not shown in the table include sample design and ACS variables.

Main Respondent Analysis

Results: Main Respondents – Data Availability

(n=3,189)	Income	Age	Race
% available	89.1	77.6	80.1
Predictors	**OR**	**OR**	**OR**
HRS Main: Financial unit structure (ref: Coupled)			
Non-coupled female	0.56*	1.31	0.75
Non-coupled male	0.49**	0.92	0.74
HRS Main: Race (ref: Other)			
Black	0.98	1.79*	1.10
Hispanic	1.41	1.95*	1.21
HRS Main: Education (ref: High School/GED)			
< High school/GED	1.08	1.13	1.06
Some college	1.06	0.73	1.25
College and above	1.10	0.74	1.14
HRS Main: Number of living child	0.85***	1.03	0.95
HRS Main: Mental health score (the higher, the worse)	0.93	1.01	0.94*
HRS Main: ADL: Some vs. No difficulty	1.47	1.43	1.09
HRS Main: Have life insurance vs. No	0.82	0.68	0.77
HRS Main: Probability of working after age 65	1.00	1.00	1.00
HRS Main: Currently working for pay vs. Not working	1.04	0.86	1.18

*Note: *p<0.05, **p<0.01, ***p<0.001*; Other predictors not shown in the table include sample design and ACS variables.

Results: Main Respondents – Data Accuracy

	Income (n=2,841)	Race (n=2,601)
% accurate	**50.6**	**79.3**
Predictors	**OR**	**OR**
PSU: Certainty vs. Non-certainty	0.75**	1.14
SSU Strata (ref: Non-minority)		
High Black	0.58**	0.34***
High Hispanic	0.62**	0.57**
High B+H	0.59**	0.27***
Address Strata (ref: LBB - Black)		
LBB - Hispanic	0.81	1.32
LBB - Other	0.59**	2.76***
Not LBB	1.15	1.33
No age	0.48**	1.51
No commercial data	0.98	1.58
Census Region (ref: Northeast)		
Midwest	1.54	0.66
South	1.41	0.73
West	1.04	0.55**

*Note: *p<0.05, **p<0.01, ***p<0.001*; Other predictors not shown in the table include ACS variables.

Results: Main Respondents – Data Accuracy (Cont'd)

	Income	Race
Predictors	**OR**	**OR**
HRS Main: Financial unit structure (ref: Coupled)		
Non-coupled female	7.64***	1.06
Non-coupled male	3.36***	1.10
HRS Main: Race (ref: Other)		
Black	1.28	2.53***
Hispanic	1.43	1.55**
HRS Main: Education (ref: High School/GED)		
< High school/GED	1.22	1.08
Some college	0.54**	1.07
College and above	0.22***	1.07
HRS Main: Number of living child	1.03	1.01
HRS Main: Mental health score (the higher, the worse)	1.12***	0.98
HRS Main: ADL: Some vs. No difficulty	1.83***	0.87
HRS Main: Have life insurance vs. No	0.68*	0.82
HRS Main: Probability of working after age 65	1.00	1.00
HRS Main: Currently working for pay vs. Not working	4.29***	1.37*

*Note: *p<0.05, **p<0.01, ***p<0.001*; Other predictors not shown in the table include ACS variables.

Implications - 1

- Commercial data is useful for signaling target age groups and Hispanics and Blacks
- Availability favors
  - Those living in areas with low minority density and outside NE (but less so for age availability)
  - Older, non-minority and coupled financial units
- Availability $\propto$ Response

Implications - 2

- Accuracy favors
  - Financial units in non-minority areas
  - Financial units currently working
  - Differential patterns between age accuracy and income accuracy
    - Income more accurate for non-coupled financial units
    - Race more accurate for minorities
- No clear sign of biases due to commercial data use, as many of the significant predictors are controlled in weighting

WORK IN PROGRESS

Thank you

Questions?

sungheel@umich.edu