Assessing Data Quality and Inference in a Web Respondent Driven Sampling Study of Ethnic Minorities

JSM, 2024

Kaidar Nurumov, Sunghee Lee

Program in Survey and Data Science
University of Michigan

Outline

1. 1. Motivation

2. 2. Health and Well-being of Koreans (HAWK)
   - Recruitment
   - Data Quality

3. 3. Implications

This work was supported by the National Institute on Aging of the National Institutes of Health [grant numbers: R21 AG062844; R01 AG060936].

1. Motivation

Data Collection for Racial/Ethnic Minorities

- Increased interest in data for granular racial/ethnic subgroups (eg. Korean Americans rather than Asian Americans)
- Uncertain feasibility of standard methods for data collection at the national level
- Web-based Respondent driven sampling (W-RDS) as an alternative
  - Racial homophily: a tendency to form social connections between individuals within the same racial/ethnic group.
    $\rightarrow$ Chain referrals in RDS
  - High web access among racial/ethnic minority groups (94.3% - ACS 2021)
  - Administration convenience on web (multiple languages, no interviewers)
- **W-RDS data quality is an open question.**

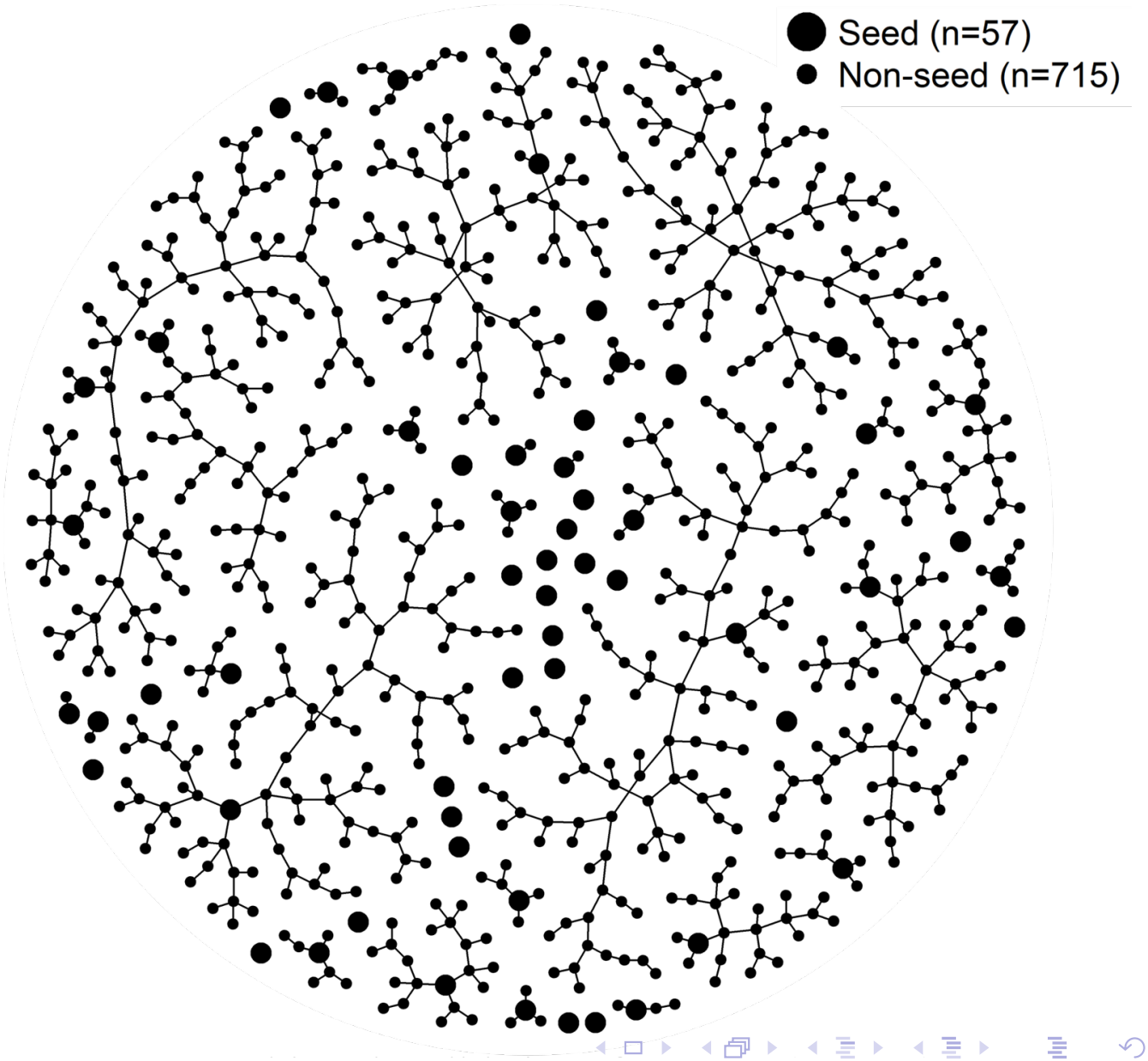
2. Health and Well-being of Koreans (HAWK)

Overview

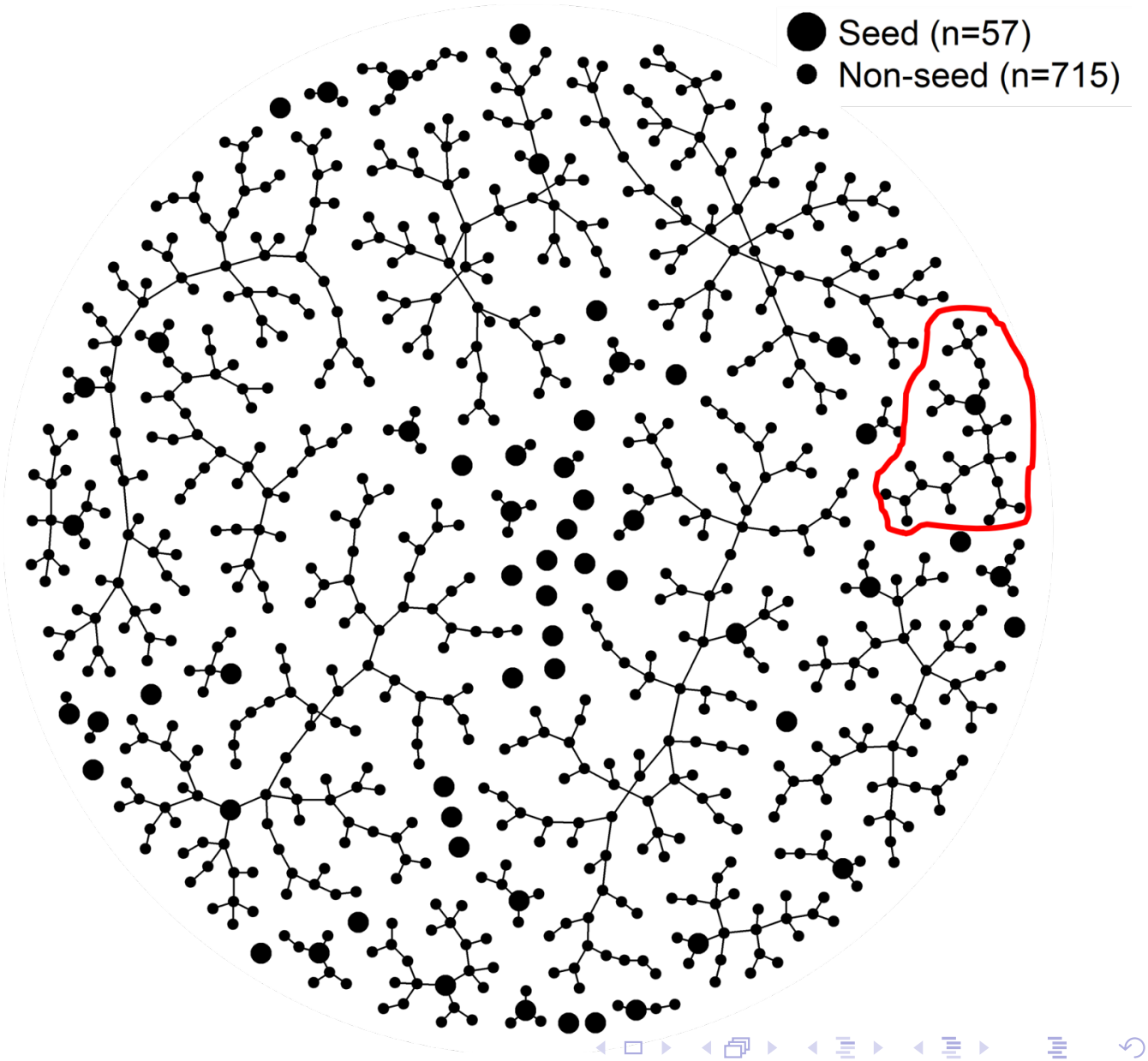
- A national survey of Korean Americans using W-RDS
- Data collection and recruitment in English and Korean
- Web questionnaire is accessed using a unique code through the study website.
- Each respondent received max. 3 recruitment coupons via email or text messages.
- Each coupon had a unique code
  $\rightarrow$ allows tracking coupon use and linking recruiters and recruits
- Incentives: \$20 for the survey; \$20 per recruit
- May 2022 - January 2023
- 57 seeds $\rightarrow n = 772$

Recruitment

Overall Recruitment Chain



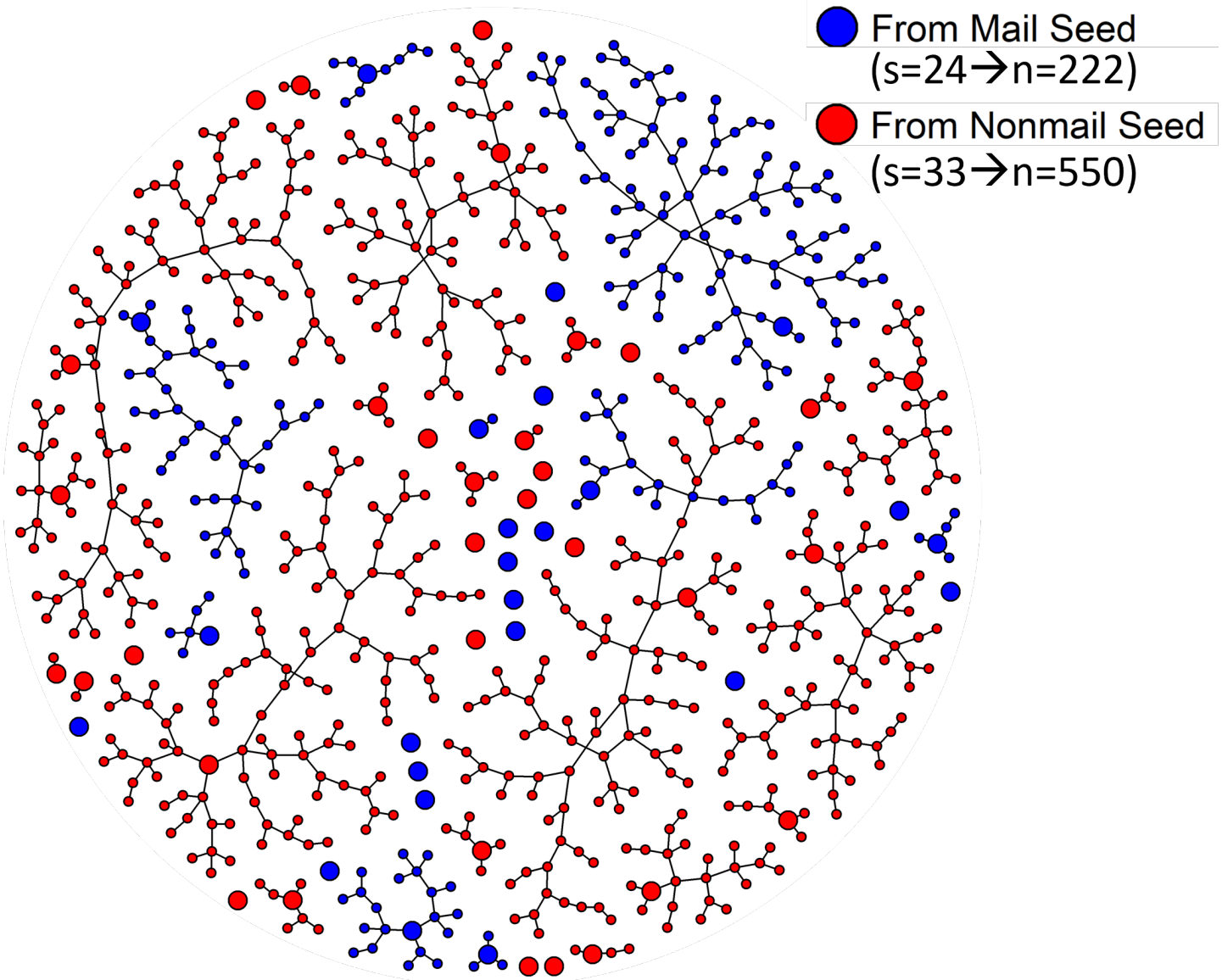
Overall Recruitment Chain



Seed Recruitment

- Mail seeds
  - List of addresses associated with Korean surnames or ethnicity
    $\rightarrow$ Mailed invitation letters with 2$ prepaid incentives
  - 24 mail seeds $\rightarrow n = 222$
- Nonmail seeds
  - Web screener through Facebook ads
    $\rightarrow$ Sent invitation emails/texts
  - 33 nonmail seeds $\rightarrow n = 550$

Recruitment Chain by Seed Type



Data Quality

Analysis Goal 1

Check data quality of overall HAWK sample

Analysis Goal 2

Check data quality of HAWK sample by seed type

Analysis Goal 3

Check data quality of HAWK sample after statistical adjustments

Methods

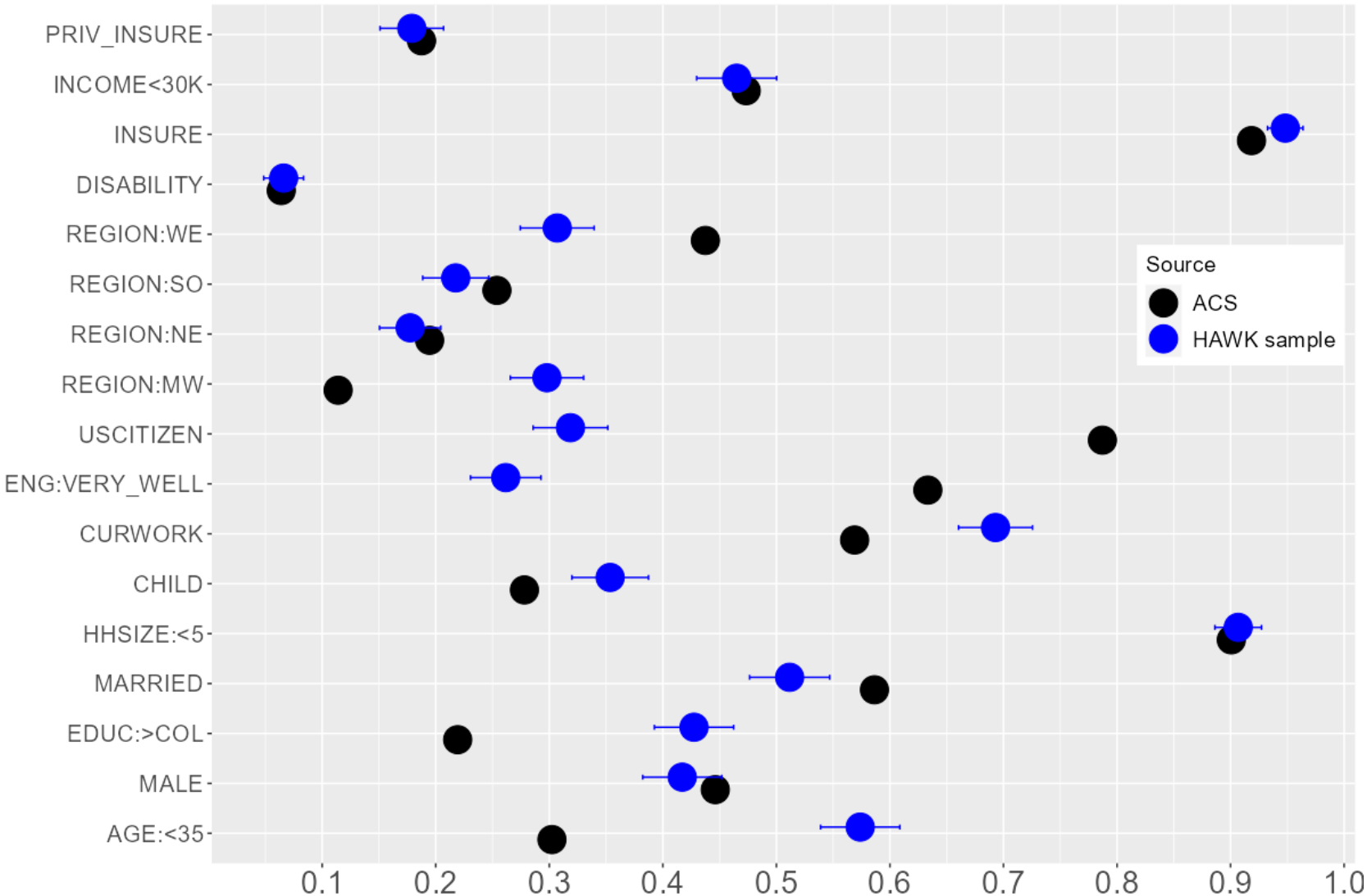
- Unadjusted HAWK
  - Assuming simple random sampling - sample estimates
- Adjusted HAWK: using Salganik-Heckathorn (RDS-I) and Volz-Heckathorn estimators (RDS-II). Salganik's bootstrap was used to calculate SEs for RDS-I and RDS-II.
- Adjusted HAWK: Propensity score
  - Create propensity scores using a weighted logistic regression
  - Common variables as predictors of being in HAWK (1) or ACS (0)
- Adjusted HAWK: Propensity scores + Raking
  - Population totals on Age, Gender, Education and Region
- American Community Survey (ACS) 2021 as a gold standard
  - Filter in Korean American adults from the sample

Adjustment applied to the overall sample, to the sample generated from mail seeds and the sample from nonmail seeds separately

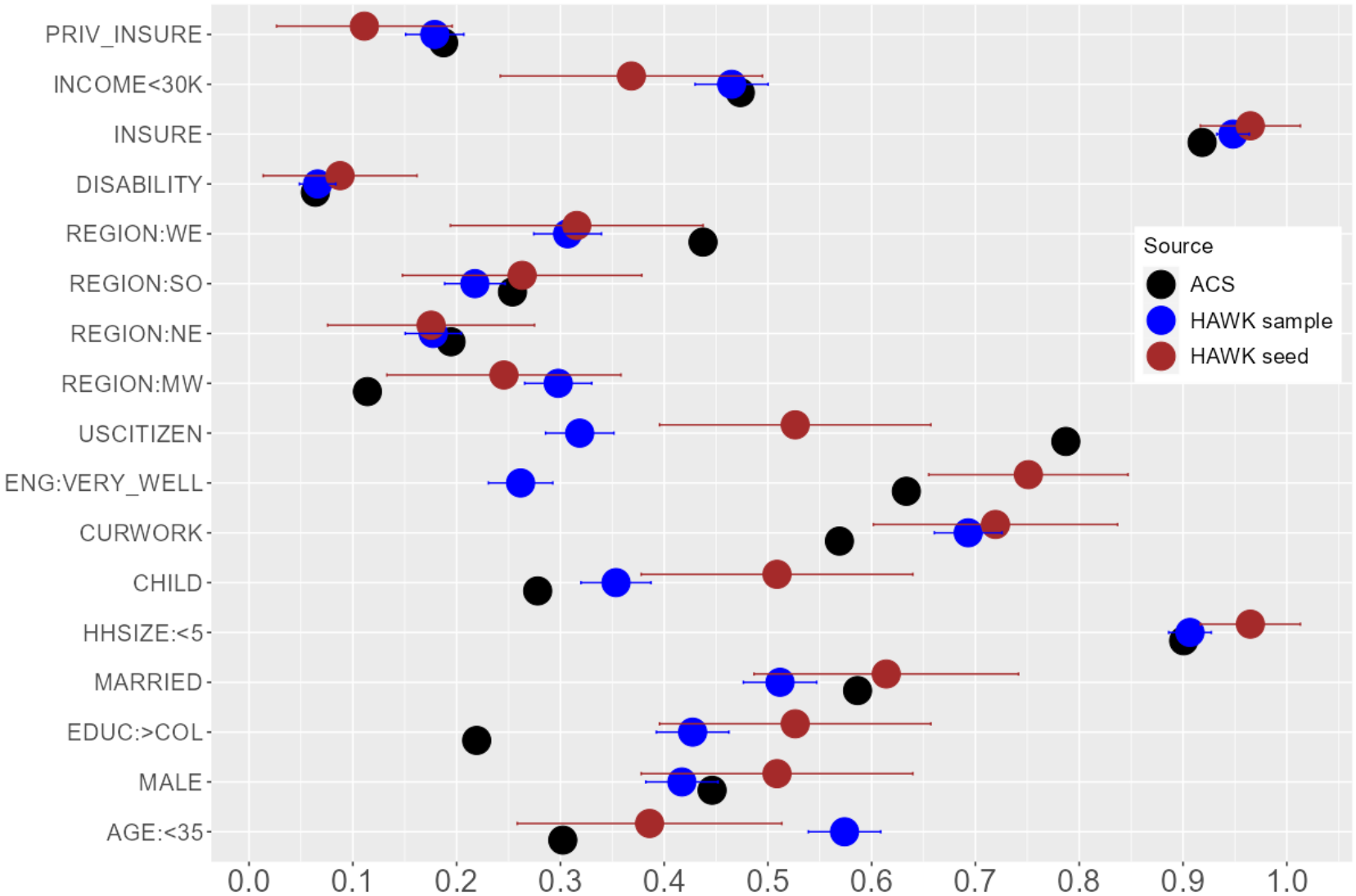
Variables Used In the Analysis

Benchmark Characteristics	Pscore	Raking
Age: % <35	X	
Sex: % Male	X	X
Educ: % >Coll.	X	X
% Married	X	
% HH members < 5	X	
% With children	X	
Employ.: % Cur. work	X	
% Spk Eng. Very Well	X	
% US citizen	X	
% Region	X	X
% Insured		
% W/ priv. ins.		
% W/ disab.		
% Inc. < 30K		

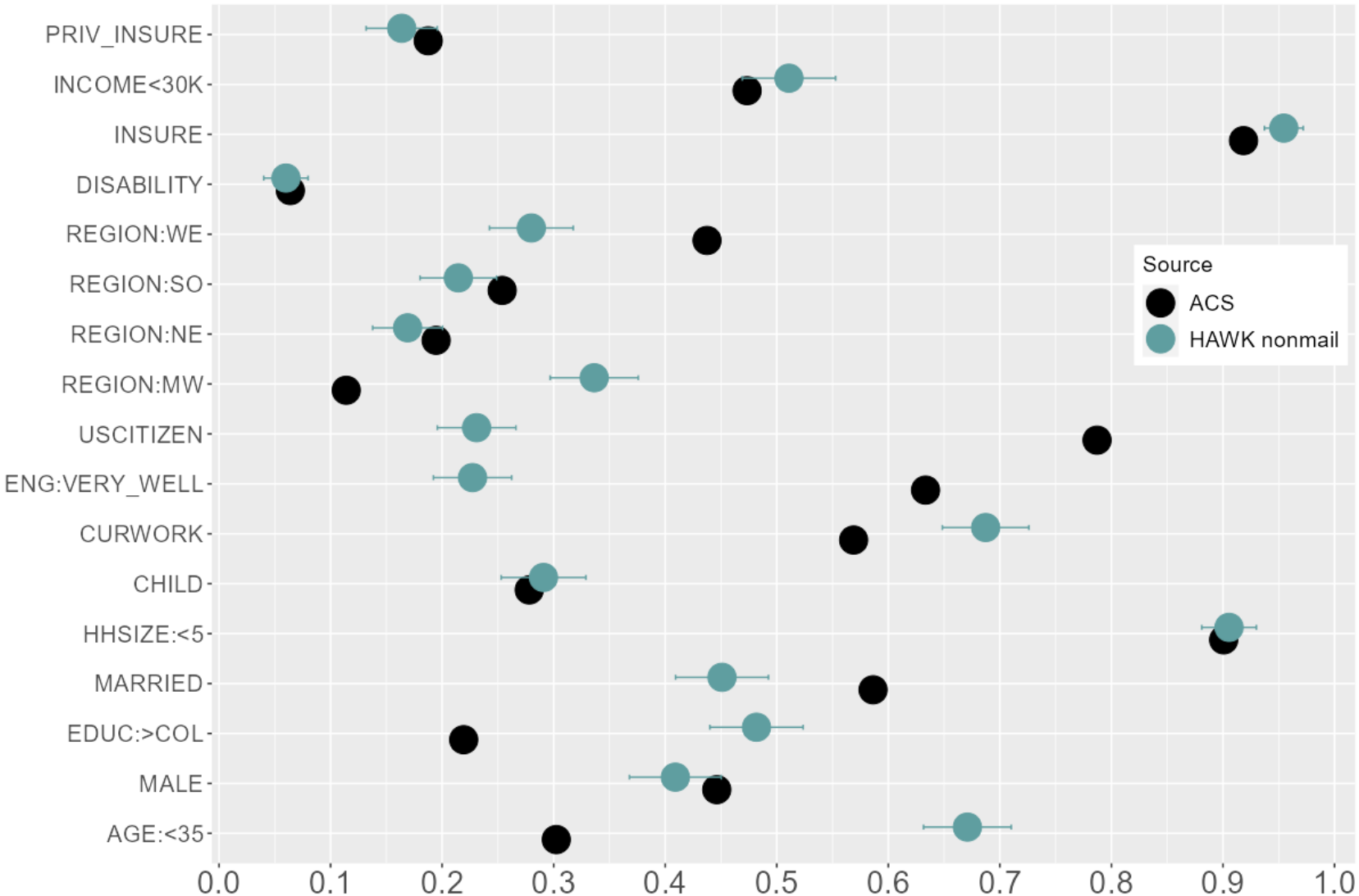
Unadj. HAWK Overall



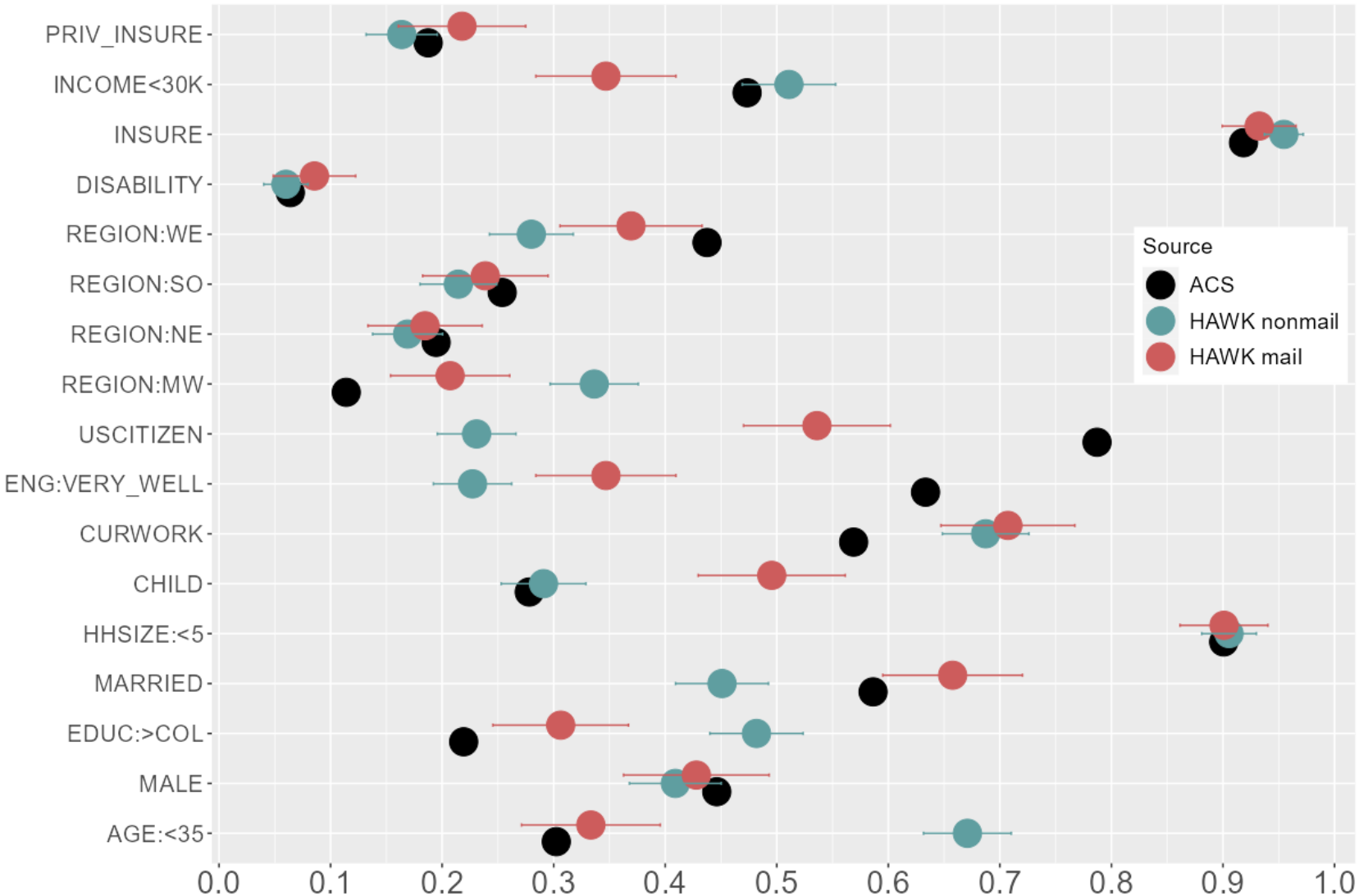
Unadj. HAWK Overall + Seed Sample



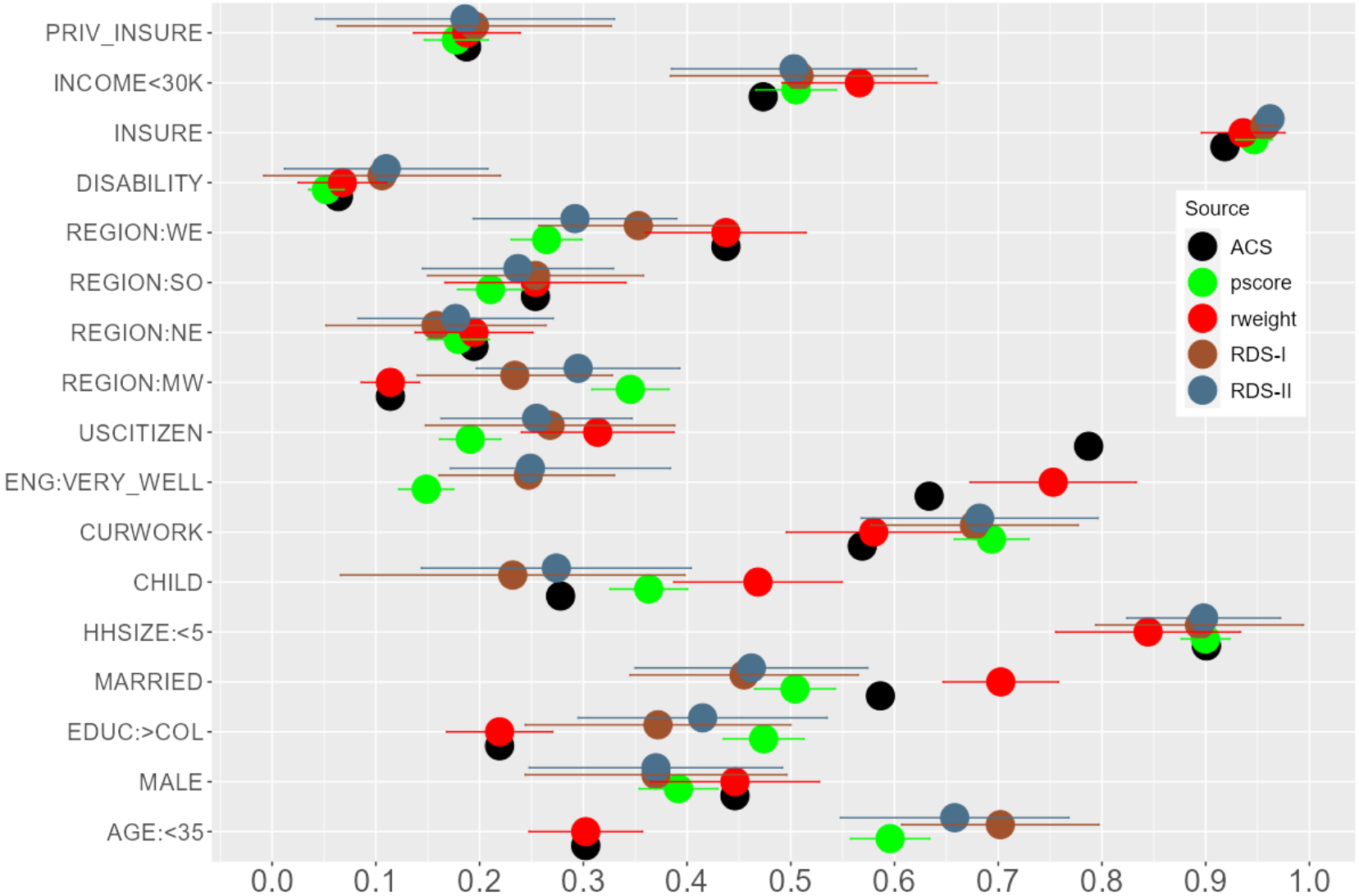
Unadj. HAWK Nonmail



Unadj. HAWK Nonmail + Mail Sample



Adj. HAWK Overall



3. Implications

Implications (preliminary)

Overall Recruitment

- 57 seeds $\rightarrow n = 772$
- 24 mail seeds $\rightarrow n = 222$
- 33 nonmail seeds $\rightarrow n = 550$

Analysis Goal 1: Data Quality of HAWK

HAWK sample (naive) is younger and more educated compared to ACS but fared well on insurance variables, disability, income, South, Northeast, HHsize and Sex.

Analysis Goal 2: Data Quality by Seed Type

HAWK sample generated from mail seeds appears closer to the Korean American population than the sample from nonmail seeds on education, age, region, US citizens, marital status, but not for income and children.

Analysis Goal 3: Data Quality after Statistical Adjustments

Adjustments do not work for some variables variables (US citizen, marital status, engl. profic.), no clear benefits.

Q & A

Thank you for your attention!

k.nurumov@umich.edu
sungheel@umich.edu