CHIS Army Knife? Examining Multiple Approaches to Sample Smaller Racial/Ethnic Groups in California

Brian M. Wells, PhD
Data Quality and Survey Methodology Manager
California Health Interview Survey
UCLA Center for Health Policy Research

JPSM/MPSDS Seminar Series – April 14, 2021

Thanks and Disclaimer

- Funding for portions of this research provided by:
  - California Rural Indian Health Board
  - California Tobacco Control Program
  - Keiro Senior HealthCare

- The analyses, interpretations, conclusions, and views expressed in this working paper are those of the author and do not necessarily represent the UCLA Center for Health Policy Research, the Regents of the University of California, or collaborating organizations or funders.

Summary

- Introduction to the California Health Interview Survey
- Highlight 4 specific examples of how CHIS attempts to improve sampling and interviewing of smaller racial/ethnic groups
  - Surname list frames and screening → Korean, Vietnamese, and Japanese
  - Secondary frames and geographic targeting → American Indians
  - Respondent-driven sampling → Native Hawaiian and Pacific Islanders
  - Predictive models using machine learning methods and third-party data
- Discussion

What is the California Health Interview Survey?

- CHIS is the nation's largest state health survey
- Most comprehensive source of health information on Californians
- Comprehensive range of health topics
  - Health status
  - Health conditions
  - Mental health
  - Oral health
  - Health behaviors
  - Access to & Use of Health Care
  - Health insurance
  - Employment
  - Respondent characteristics

What is the California Health Interview Survey?

- California's assessment tool to meet state and local needs for population-based health data
  - Policy analysis, development, and advocacy
  - Service and program planning
  - Research
- CHIS is a collaborative project, funded by federal and state health agencies, California and national foundations, and others

CHIS Design Over the Years

- Conducted every other year since 2001, annually since 2011
- Up to 3 household interviews
  - An adult (age 18+) in the household, adolescent (ages 12–17) if present, and child (ages 0–11) if present
- Over 40,000 adult interviews per cycle (over 2 year cycle since 2011)
- Conducted in multiple languages
  - English, Spanish, Chinese*, Korean, Vietnamese, and Tagalog

    * both Cantonese and Mandarin dialects

CHIS Design Over the Years

CHIS Cycle	CHIS 2001	CHIS 2003	CHIS 2005	CHIS 2007	CHIS 2009	CHIS 2011-2012	CHIS 2013-2014	CHIS 2015-2016	CHIS 2017-2018	CHIS 2019-2020	CHIS 2021-2022
Sampling frame	Landline RDD				Dual-frame landline and cell RDD (80/20)			Dual-frame landline and cell RDD (50/50)		Address-based sampling	
Data collection mode	CATI									Web survey w/ CATI nonresponse follow-up	
Contact mode	CATI w/ advance letter to phone numbers linked to addresses									Multiple mail contacts w/ CATI follow-up to addresses linked to phone number	

Sampling Goals of CHIS

- The sample is designed and optimized to meet two objectives:

1. Provide estimate for large- and medium-sized counties in the state, and for groups of the smallest counties (based on population size)
2. Provide statewide estimates for California's overall population, its major racial and ethnic groups, as well as several racial and ethnic subgroups

- How does CHIS accomplish goal #2 for different racial and ethnic groups?

#1: Oversampling Korean, Vietnamese, and Japanese

Asian Subgroups in California

- American Community Survey (ACS) estimates over 5.8 million Asians in California in 2019, or about 15% of the state's population
  - Majority are Chinese (~1.9 million), Filipino (~1.5 million), and Asian Indian (~800,000)
  - ~700,000 Vietnamese (4th largest Asian subgroup)
  - ~500,000 Korean (5th largest Asian subgroup)
  - ~350,000 Japanese (6th largest Asian subgroup)
- Strong geographic clustering in some cities, zip codes

Ethnic Surname List Frames

- Surname lists produced from multiple sources (Social Security Administration, Census, etc.) corresponding to Spanish, Chinese, Japanese, Filipino, Korean, Asian Indian, and Vietnamese
  - For example, Vietnamese surnames: Nguyen, Tran, Le, and Pham
- Surname flags from third-party data can be merged with other sampling frames (e.g., ABS)
- Surname list frames then be used for targeted sampling or disproportionate stratified sampling

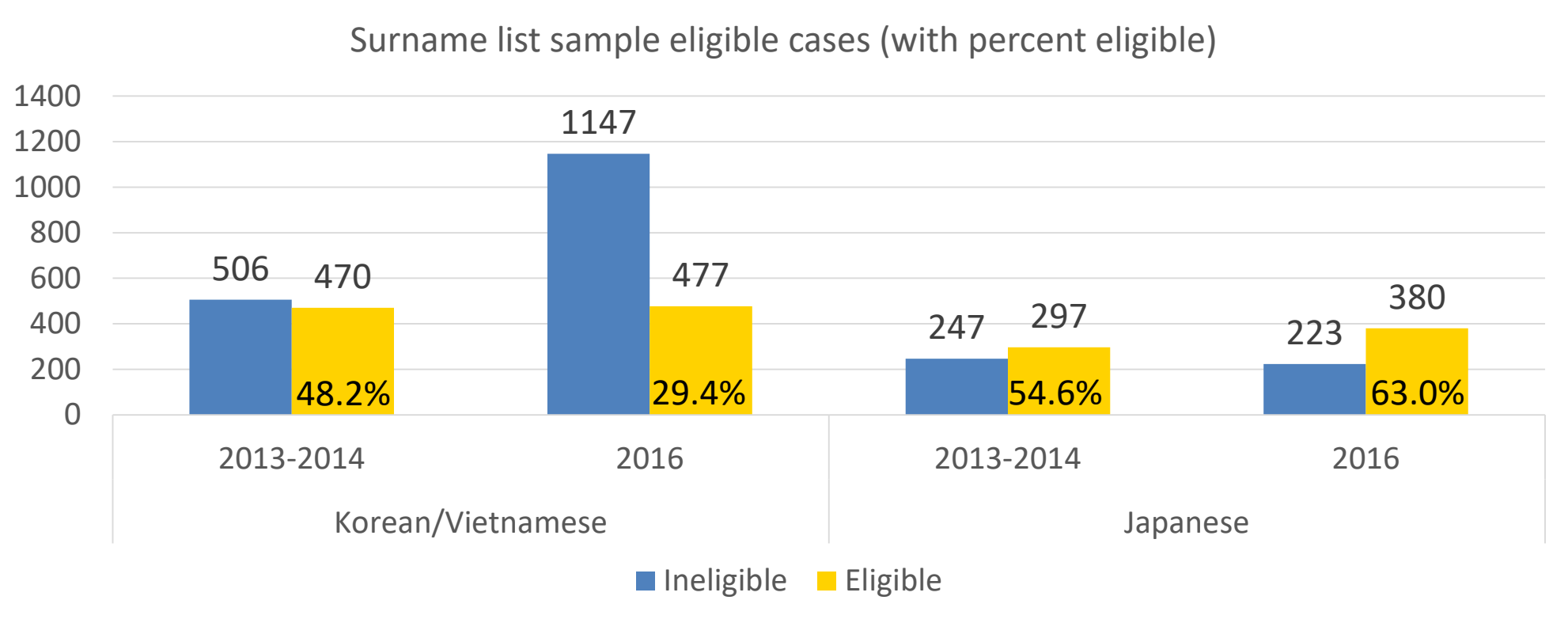
e.g., Lauderdale & Kestenbaum, 2000; Elliott et al., 2009; Shah et al., 2010; Taylor et al., 2011

Disproportionate Stratified Sampling and Screening

- CHIS 2003-2018 used disproportionate stratified sample from Korean and Vietnamese surname list frames
- CHIS 2003-2014 screened for Korean and Vietnamese Californians
  - "Do any of these adults who live in your household consider themselves to be Korean or Vietnamese or of Korean or Vietnamese descent?"
  - Screener removed for 2015-2018
- Similar approach in CHIS 2014-2016 for Japanese Californians using Japanese surname list frame, and a screening question in 2014
- What is the impact of removing the screening?

Accuracy of Ethnic Surname List Frames in CHIS

- 2013-2014 screener vs. 2016 completed interviews
  - Not equivalent measurements; relative comparison



Source: California Health Interview Survey, 2016

Eligible vs. Ineligible Comparison

- About 70% of ineligible Korean/Vietnamese cases in 2016 completed an adult interview
  - 58% identified as Chinese with 14% completing in Mandarin or Cantonese
  - More likely to be US citizens, college graduates, proficient English speakers
- 37% ineligible for Japanese cases in 2016
  - 86% identified as non-Asian
  - More likely to be non-citizens, low education, low English language proficiency

- For more details, see Wells et al. (2018) and Becker et al. (2018)

#2: Oversampling American Indian and Alaska Natives

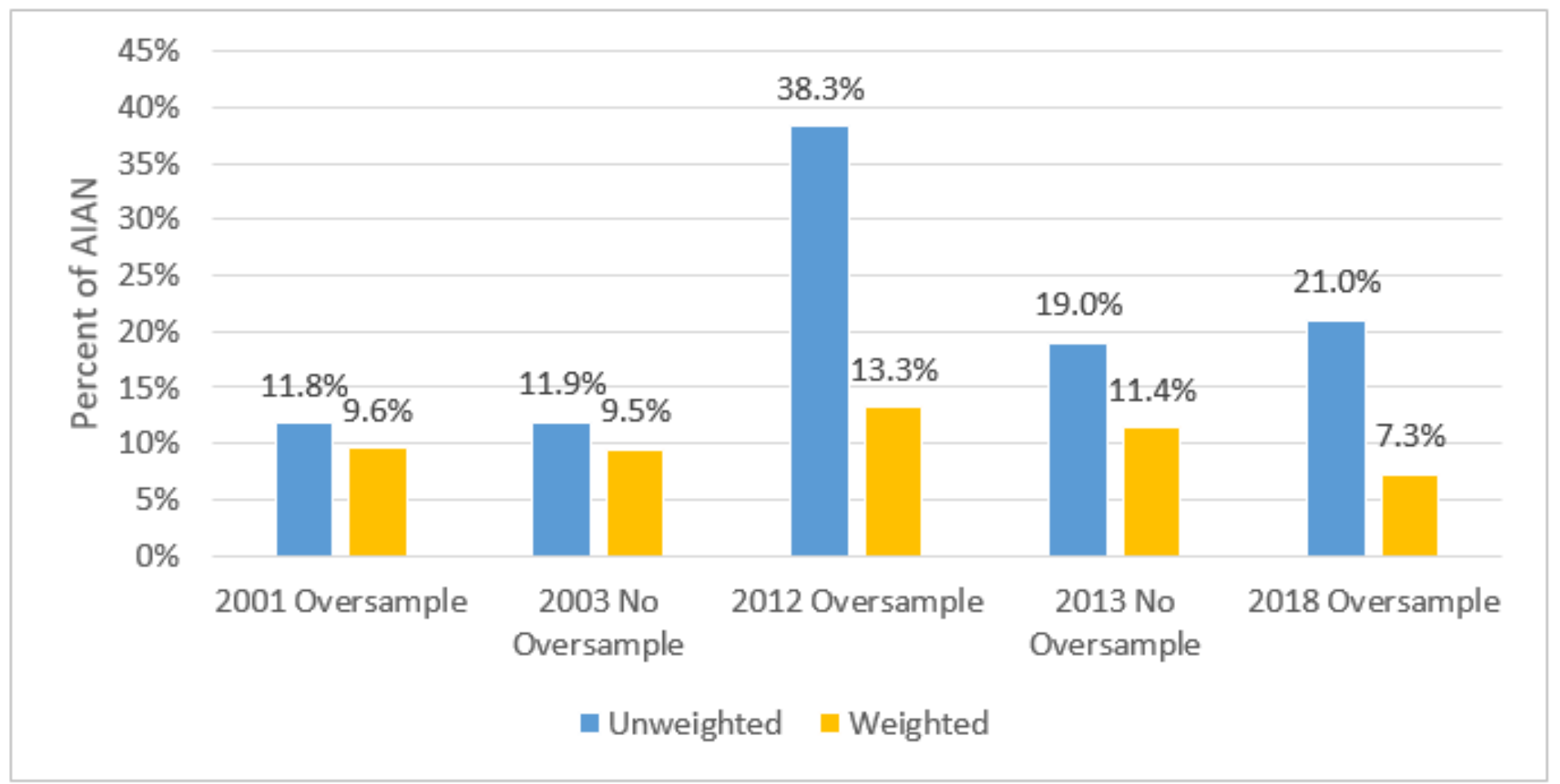
American Indians in California

- Over 770,000 American Indian and Alaska Natives (AIAN) in California
- Differences in counts based on self-identification vs. tribal affiliation
  - Over 300,000 are a member of a federally recognized tribe
  - Only about 100,000 have a California tribal affiliation
- Tribal membership allows for benefits like access to Indian Health Services (IHS)
- In and around reservations are common places to find AIAN
  - Harder to find urban AIAN

Different years, different approaches

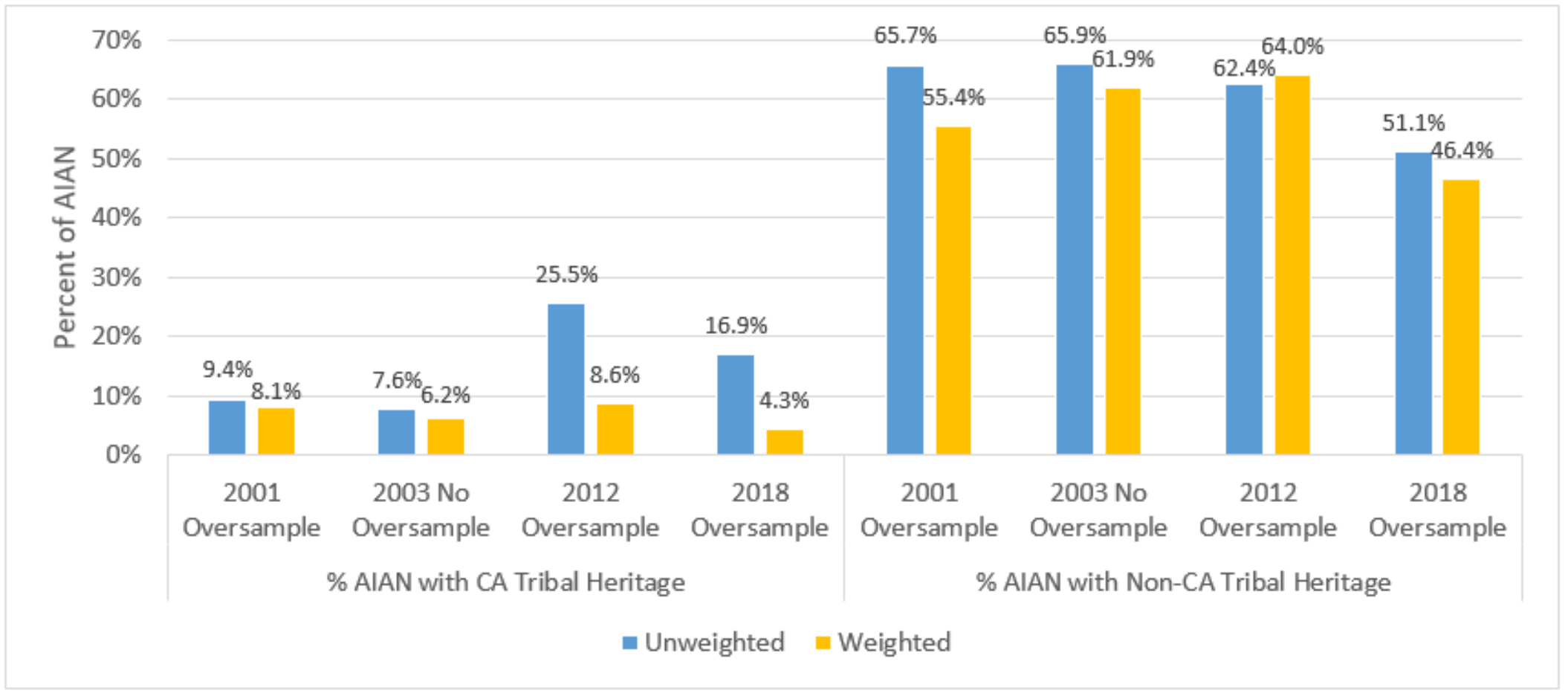
- CHIS 2001
  - Sampling list developed with input from AIAN tribal organizations, an Indian Health Services (IHS) clinic user list, and stratification by urban/rural status
  - IHS eligibility is based on membership in federally recognized tribes
- CHIS 2012
  - Sampling list based on IHS clinic user list
- CHIS 2018
  - IHS clinic user list unavailable
  - Disproportionate stratification using listed frame with AIAN flag
  - Telephone-matched addresses in medium and high incidence Census blocks

Percent AIAN Enrolled in a Recognized Tribe



Source: California Health Interview Survey, 2001, 2003, 2012, 2013, 2018

Percent AIAN with California Tribal Heritage



Source: California Health Interview Survey, 2001, 2003, 2012, 2018

#3: Oversampling Native Hawaiian and Pacific Islanders

Native Hawaiian and Pacific Islanders in California

- 335,000 Native Hawaiian and Pacific Islanders (NHPI) within state
  - ~75,000 Native Hawaiian
  - ~60,000 Samoan
  - ~45,000 Guamanian and Chamorro
- Strong geographic clustering around the Los Angeles metropolitan area and San Francisco Bay area
- Tight-knit, family-orientated communities
- Extended families distributed across the state suggest these geographic communities are linked, are not independent

Respondent-Driven Sampling (RDS)

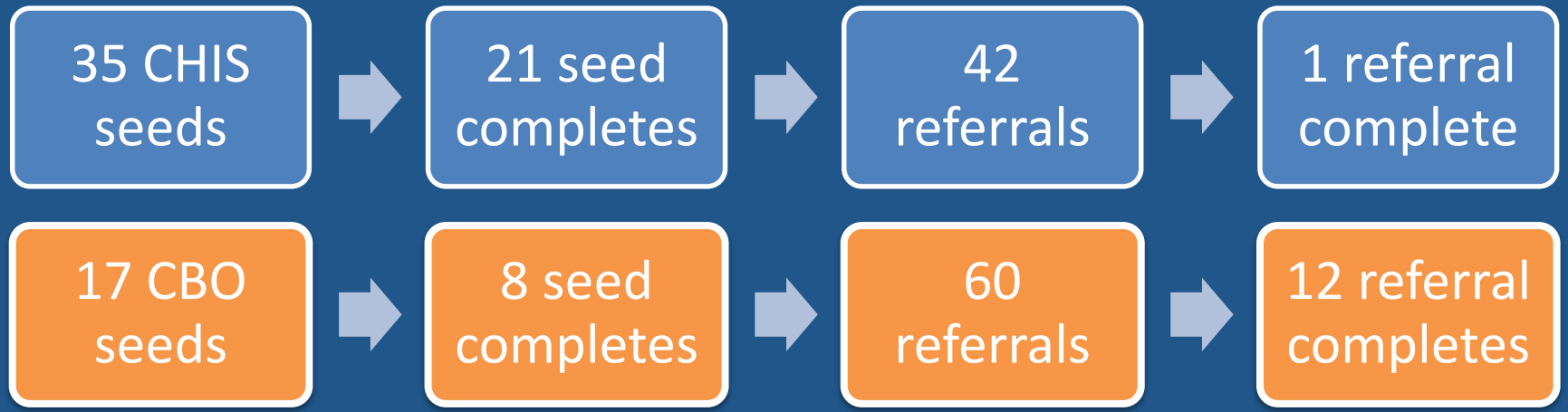
- RDS is an extension of snowball sampling where seeds are purposively selected and participants are incentivized to recruit their eligible peers (Heckathorn, 1997)
- Obtain social network size information to adjust sampling weights to approximate a probability sample
- Often used for hidden populations (e.g., drug users, people with HIV)
- Historically conducted through in-person methods, but recent growth into web-based respondent-driven sampling (Web-RDS)
- For example, Web-RDS has been successful for Koreans in Los Angeles and Michigan (Lee et al., 2020)

Applying RDS for NHPI in California

- Seeds from two sources
  - CHIS respondents (identifies as NHPI, willing to participate in follow-up)
  - Strong advocates from NHPI community-based organizations (CBOs)
    - Maximum of 2 seeds per CBO
- Incentive structure
  - $20 to complete a 15 minute web (or CATI) survey
  - $10 per successful referral (maximum of 3)
- Send up to 3 referral coupons by their preferred method: email, text, or physical copy (mail)
- Not part of the main CHIS study; scheduled to field in Summer 2020

NHPI RDS Results

- 42 completes, 29 of which were seeds (69%)



Re-scoping Amidst the COVID-19 Pandemic

- Study-specific Web-RDS methods did not seem to work with how NHPI communicate and network
  - COVID-19 pandemic another complication
- Stronger emphasis needed geared toward specific NHPI communities and how NHPI communicate (e.g., Facebook)
- To achieve desired sample goals, study needed to transition to a convenience sample
- Survey advertised by CBOs through their email and social media networks, $20 incentive
- Currently in the field, too early to share results

#4: Machine Learning Methods for Predictive Modeling

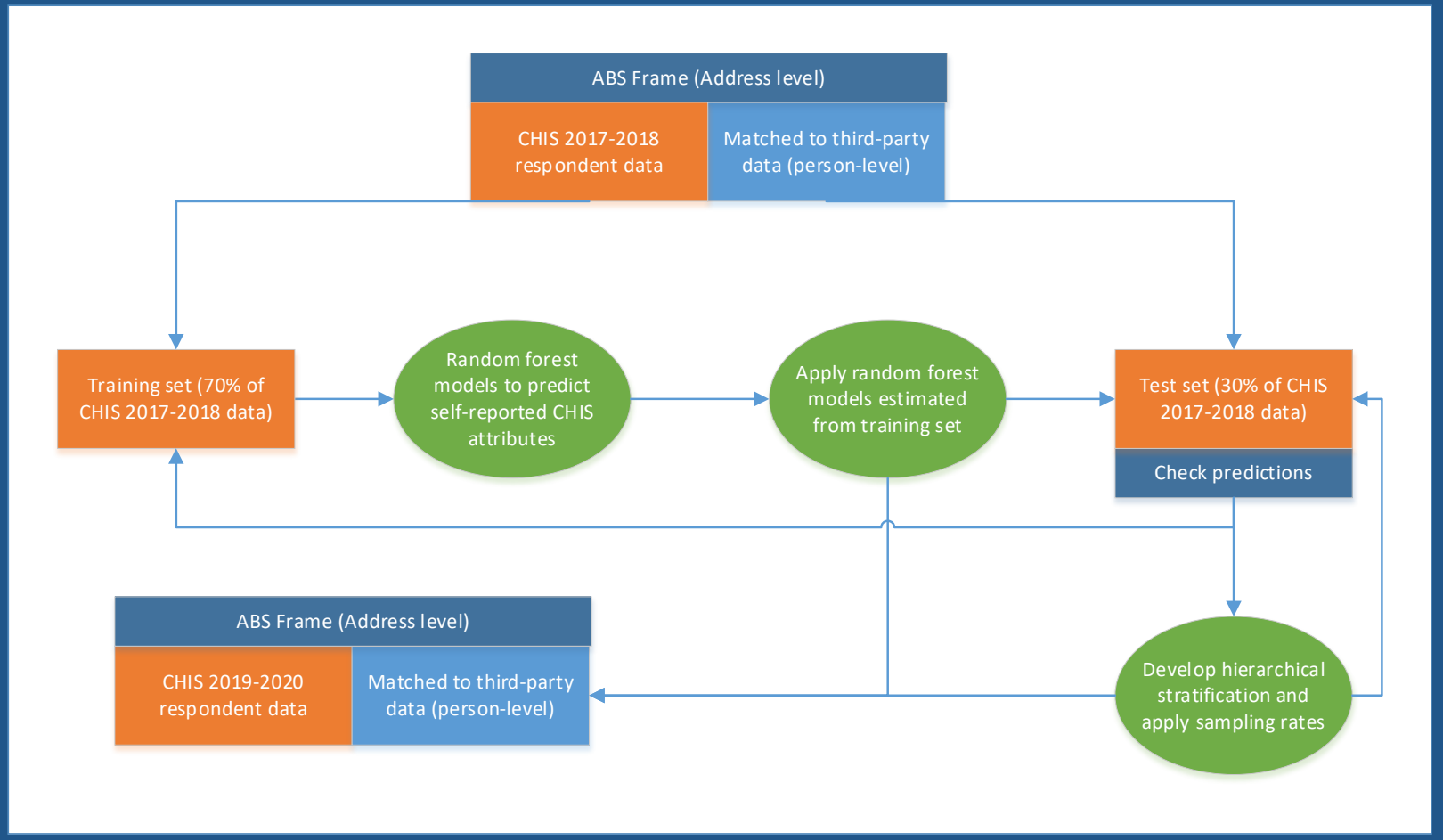
Appending External Data Sources to Stratify Sample

- ABS frame can be appended with third-party consumer databases, voter registration, etc. to allow for additional stratification to target needed or underrepresented groups
  - Age-related household flags helped improve sampling efficiency in a national fertility study (West et al., 2015)
  - Households with young children (Barron et al., 2015)
- Quality of third-party data sources unknown
- Apply machine learning approaches with third-party in conjunction with self-reported survey data to predict attributes (e.g., Dutwin, 2018)

Predictive Modeling for CHIS

- CHIS 2017-2018 used to “build” models predicting key attributes/ demographics
  - Independent (predictor) variables appended from multiple data sources
    - Voter registration & consumer databases
    - Sample vendor auxiliary information (e.g., surname indicators)
    - Geo-demographic data from Census/ACS (e.g., proportion of Hispanics by block group)
- Most effective models retained and used to assign predictive probability score to every sampled address in CHIS 2019-2020

Appending External Data Sources to Frame Data



Predictive Modeling and Sampling Goals

- CHIS strives to provide estimates for major racial and ethnic groups, as well as several subgroups
- Based on sampling goals, shortfalls observed in the Fall 2018 Pilot, and model efficiency, CHIS 2019 used the following sampling strata:
  1. Korean
  2. Vietnamese
  3. Other Asian
  4. Hispanic and Spanish-speaking
  5. Low educational attainment
  6. Non-U.S. Citizens
  7. Households with children/teens
- Each sampled address assigned to sampling strata hierarchically
  - Strata sampled disproportionately to ensure representation of key subgroups

Predictive Modeling Results

Characteristic/Strata	Total number of Interviews	Number of interviews from targeted stratum	Percent of interviews from target stratum	Estimated incidence without oversampling	Incidence in targeted stratum	Overall achieved incidence
Korean	625	237	37.9%	1.3%	39.0%	1.4%
Vietnamese	509	292	57.4%	1.1%	47.9%	1.2%
Other Asian	4,592	1,028	22.4%	10.7%	26.9%	10.5%
Hispanic/ Interview conducted in Spanish	8,203	2,111	25.7%	18.2%	55.2%	18.7%
Less than High School or Non-Citizen	3,068	263	8.6%	7.5%	10.5%	7.0%
Child or Teen*	6,652	2,436	36.6%	13.2%	31.2%	15.2%

*Note: Numbers presented here are based on completed child/teen interviews in CHIS 2019-2020

Predictive Modeling Results (continued)

- Target group incidence in target strata indicate generally successful models
  - Allowed oversampling of harder-to-reach subgroups (e.g., Hispanics)
  - Allowed targeting of very low-incidence subgroups (e.g., Vietnamese & Korean)
- Some models performed better than others
- Future oversampling fractions can be boosted to account for incidence as well as response rates within strata (Hispanic and lower education strata have significantly lower response rates)
- Stratification needs to account for overlap within strata to lead to most efficient design

Discussion

Some Final Thoughts...

- Each population of interest requires different approaches or may have different tools available for sampling
- Ethnic surname flags are reasonable for Asian subgroups, but are not available for other groups like AIAN, NHPI, but can work for Hispanic surnames
  - Historically a relatively efficient tool
- IHS service frame is a unique resource for the AIAN, but most other race/ethnic populations do not have such a list
  - Third-party flags not effective for AIAN

Some More Final Thoughts...

- Respondent-driven sampling for sampling racial/ethnic group requires additional work
  - Researchers need to have a solid understanding of their population, how they connect and communicate
- Machine learning methods with third-party data have potential, but the success is dependent on the sensitivity and specificity of the models, as well as the quality of the indicators
  - Model for Asian subgroups and households children good
  - Preliminary models for AIAN or NHPI were inefficient

Thank you!

Brian M. Wells, PhD

bmwells@ucla.edu

References

- Barron, M. et al. (2015). Using auxiliary sample frame information for optimum sampling of rare populations. *Journal of Official Statistics,* 31(4), 545-557.
- Becker, T., Wells, B. M., & Ponce, N. (2018, May). *A respondent by any other name: The impact of interviewing respondents sampled from ethnic surname phone lists who don't meet racial/ethnic criteria*. Paper presented at the 2018 American Association for Public Opinion Research Conference, Denver, CO.
- Dutwin, D. (2018, Oct.). *Feedback loop: Using surveys to build and assess registration-based sample religious flags for survey research.* Paper presented at the BigSurv18 Conference, Barcelona, Spain.
- Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*(2), 174-199.
- Lauderdale, D. S. & Kestenbaum, B. (2000). Asian American ethnic identification by surname. *Population Research and Policy Review, 19*(3), 283-300.
- Lee et al. (2020). Respondent driven sampling for immigrant populations: A health survey of foreign-born Korean Americans. *Journal of Immigrant and Minority Health.*
- Wells, B. M., Becker, T., & Ponce, N. (2018, May). *Increased racial and ethnic diversity from using surname list samples in the California Health Interview Survey*. Paper presented at the 2018 American Association for Public Opinion Research Conference, Denver, CO.
- West, B. T., Wagner, J., Hubbard, F., & Gu, H. (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*(2), 240-264.