# Oversampling multigenerational families & immigrant groups in the NJ Population Health Cohort Study

Katherine Morton[1], Steven Cohen[1], Stephanie Zimmer[1] and Joel C. Cantor[2]
[1]RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194
[2]Institute for Health, Health Care Policy, and Aging Research, Rutgers University

**Abstract**
The New Jersey Population Health Cohort Study is the largest study to date to explore factors that influence health and well-being in the state. More specifically, the study seeks to improve our understanding of how life events and stress affect health and resilience, and identify health disparities encountered by historically disadvantaged groups, multigenerational families, and immigrant groups. To help ensure that precision targets for planned analytical investigations are achieved for the overall study and specifically for multigenerational families and immigrant groups, oversampling strategies are employed. The study also includes some nonprobability-based design features that include respondent driven sampling to include rare populations of immigrants. In this manuscript, attention is given to the study's analytic objectives, features of its sample design, its implemented oversampling strategies, and expected sample yields. Planned analyses will examine the performance of the sampling strategy for identifying groups of interest.

**Key Words:** NJ Health Cohort Study; health disparities; multigenerational families; immigrant groups; model-based oversampling; Address-Based Sampling (ABS)

## 1. Background

The New Jersey Population Health Cohort Study (NJHealth) is the largest study to date to explore factors that influence health and well-being in the state. Led by the Rutgers Institute for Health, Health Care Policy and Aging Research, the study seeks to improve our understanding of how life events and stress affect health and resilience and to generate theoretically grounded and actionable knowledge to improve health and well-being in the population overall and among diverse groups with a high likelihood of exposure to stressors over the life course (Cantor et al. 2023). The study was designed to enroll and follow longitudinally up to 10,000 participants from across New Jersey, with an emphasis on historically disadvantaged groups, multigenerational families, and immigrant groups, including those from China, Dominican Republic, Haiti, India, Jamaica, Korea, Mexico, Nigeria, or the Philippines and those entering the US seeking asylum, under temporary protected status, or related immigration authorities. Specific aims of the NJHealth Study are to a) identify how enduring and emerging stressors over the life course contribute to health in diverse populations, and b) discover novel factors that buffer or amplify these influences on personal and population health.

In this manuscript, attention is given to the study's multi-stage probability sample design, the use of address-based sampling (ABS) and the model-based oversampling strategies utilized to help to increase the yields and precision of survey estimates of recent immigrant populations in the state and of multigenerational households. To help ensure that precision targets for planned analytical investigations are achieved for the overall study and specifically for multigenerational families and immigrant groups, oversampling strategies are employed. The study also includes some nonprobability-based design features that adapt respondent driven sampling to include rare populations of immigrants (Gile et al 2018; Handcock 2017; Heckathorn, 1997). In this manuscript, attention is given to the study's analytic objectives, features of its sample design, its implemented oversampling strategies, and expected sample yields.

## 2. Overview of the Probability Sample Design

The NJHealth Study has a goal of producing state-wide estimates representative of the household population with an emphasis on 1) oversampling recent immigrants and 2) oversampling multigenerational households. Additionally, there is a goal of ensuring adequate sample sizes for racial/ethnic minority and low-income groups. The goals of the study are achieved through a multi-stage probability sampling design.

The probability-based sampling design explicitly oversamples recent immigrants and multigenerational households at distinct stages and will oversample Hispanic and Black populations, as shown in Table 1. In addition, the design is expected to represent the Asian population and households in the lower income quintile (household income less than or equal to $27,800) in alignment with their proportional representation in New Jersey.

Table 1. Comparison of Population and Estimated Sample Percentages for Select Domains

| Domain | Population Percentage* | Estimated Sample Percentage |
|---|---|---|
| Recent Immigrants | 4.5 | 5.1 |
| Hispanic | 20.4 | 22.8 |
| Black | 12.7 | 14.4 |
| Asian | 9.5 | 9.7 |
| Low Income Household | 20.0 | 20.2 |

*Source: 2015-2019 5-year American Community Survey.

A total of 96,552 housing units will be selected for the study. Assuming a 5% response rate, approximately 4,800 responding households are anticipated for the probability-based sample, yielding 6,000 completed survey interviews at the individual level.

Overall, the sample design implements a four-stage probability sampling approach to select N=6,000 individuals living in New Jersey households. In addition to being designed to represent the State's households overall, the design oversamples multigenerational and low-income families and non-Hispanic Black and Hispanic individuals. We use a clustered, address-based sample (ABS) to enable efficient in-person data collection. Sampling is performed by RTI International using its augmented ABS sampling frame (American Association for Public Opinion Research 2016; Harter et al 2018; Harter & McMichael 2012; McMichael & Wiant 2019; Shook-Sa 2013).

**Stage 1 Sampling: Primary Sampling Units (PSUs) – (PUMAs).** Public use microdata areas (PUMAs) are sub-state geographic units created by the Census Bureau. They are large enough to have microdata from the American Community Survey (ACS) published for each area and each have at least 100,000 people. New Jersey has 73 PUMAs in the state. The primary sampling units (PSUs) for this survey are constructed from these PUMAs. Most PSUs consist of only 1 PUMA with an exception being for Jersey City and Newark City, which are each comprised of 2 PUMAs.

The PUMAs are sampled proportional to size where the size is defined as:

$$MOS_i = PopNonImm_i + 3 \times PopImm_i$$

where $PopNonImm_i$ is the population which are not recent immigrants in PSU $i$ and $PopImm_i$ is the population which are recent immigrants in PSU $i$ where recent immigrants are defined as the

population entering the US since 2010 according to the 2015-2019 ACS[1]. This specification oversamples areas with more recent immigrants.

### 3. Scheduling PSU releases

The original design specified that interviewing would occur in 3 distinct waves with sample included in a representative subset of the PSUs each wave. The waves are characterized by an unequal number of PSUs released for each of the data collection operational periods to help best utilize the available field staff in an efficient and effective manner. There are 8, 10, and 12 PSUs released in waves 1, 2, and 3, respectively. Samples are released via representative subsets of PSUs to take into account field staff capacity and travel requirements. Consequently, each wave serves as a representative sample of the targeted study population in New Jersey, though the variances of wave specific survey estimates will be substantially larger than those achieved for the full study sample. As each wave-specific sample is added, the variance of the pooled estimates will be noticeably reduced. The attraction of the design to permit wave-specific estimates of the target population in New Jersey is the acquisition of fast-track survey estimates. This model allows for survey results to be analyzed and published for earlier waves and not require the entire data collection to be completed prior to learning of the study findings.

### 4. Subsequent Stages of Sample Selection

**Stage 2 Sampling: Second Stage Units (SSUs) - (CBGs).** Census block groups (CBGs) are the smallest area that Census publishes 5-year ACS estimates. In New Jersey, there are 6,320 CBGs with an average of 572 housing units and median of 506.5 housing units ranging from 0 to 4,347 housing units. Using an internally developed algorithm, secondary sampling units (SSUs) are constructed from CBGs by collapsing neighboring CBGs within each PUMA as necessary until all SSUs have at least 300 housing units as 200 housing units are later sampled within each SSU. After constructing the SSUs, a random sample of 23 SSUs is selected from each PSU. These are selected proportional to size where the size is defined as:

$$MOS_{i,j} = PopNonImm_{i,j} + 3 \times PopImm_{i,j}$$

where $PopNonImm_{i,j}$ is the population who are not recent immigrants in SSU $j$ in PSU $i$ and $PopImm_{i,j}$ is the population who are recent immigrants in SSU $j$ in PSU $i$. In total, 690 SSUs are sampled.

**Stage 3 Sampling: Housing Units.** Within each SSU, the housing unit frame was constructed using addresses from an address-based sampling (ABS) frame derived from the U.S. Postal Service's Computerized Delivery Sequence file. Addresses on the ABS frame are assigned to SSUs through a process called geocoding. Geocoding involves assigning a location (latitude and longitude) to an address by comparing the descriptive location elements in the address to those present in geocoding locator files. Addresses are geocoded using ESRI's Business Analyst USA Composite locator which is comprised of several individual locators. The input addresses are first matched against the locator with the highest geographic accuracy which is the rooftop location of addresses. If a suitable match is not found, those unmatched input addresses cascade to the next locator, which interpolates the position along a street using the address range. This process continues through the locators, with each one returning decreasing geographic accuracy. If no match can be found for the last locator, the input address is returned as unmatched.

Within each SSU, an initial sample size of 200 housing units is selected. Using survey data and address data from a prior study, a model was built to predict whether households are multigenerational. The survey used in modeling was a nationally representative sample which

[1] All ACS data in this report is obtained from the 2015-2019 5-year American Community Survey. At the time the sample design was prepared, the 2020 ACS data had not been released.

included questions on the ages of the residents of the household. Using these ages, a multigenerational indicator was derived and used as the outcome in the model. Data used in the model as covariates came from a proprietary marketing database which has data on most addresses in the United States and area level data from the ACS. The final logistic model consisted of the following terms:

- Multigenerational indicator based on Age C (Marketing)
- Person aged 40-59 age B (Marketing)
- Average number of persons per HU (ACS)
- Young adult in household (Marketing)
- Number of children in household (Marketing)
- Percent population aged 18-24 (ACS)
- Percent population white, NH (ACS)
- Percent of HUs where householder moved to current unit in 2010 or later (ACS)
- Percent population white, NH (ACS) interacted with Multigenerational indicator based on Age C (Marketing)
- Percent of HUs where householder moved to current unit in 2010 or later (ACS) interacted with person aged 40-59 age B (Marketing)
- Any adult age 55-64 (Marketing)

On the validation data, the area under the curve (AUC) was 70%. The AUC measures the ability of a classifier to distinguish between classes with an AUC being a value between 0 and 100% with an AUC of 50% indicating the model does no better than a coin toss and AUCs greater than 50% indicating the classifier can distinguish between classes with better accuracy than a coin flip. To help maximize the correct classification rate, a threshold of 48.2% was used. That is, if a prediction is 48.2% or greater, then the housing unit is assigned to be in the strata that is more likely to have multigenerational members, namely the high-density stratum. On the validation data, using this threshold, the prediction is correct 74.2% of the time.

Addresses are oversampled in the high-density stratum. In wave 1, 8.4% of addresses in the selected SSUs were in the high-density stratum but the sample was allocated so 19.3% of the addresses selected were from the high-density stratum (Zimmer et al 2022). Overall, the initial sample planned for Wave 1 had 38,694 cases, subsequently reduced to 25,928 cases. Wave 2 and Wave 3 will be selected at a later date as the household oversampling rate may change.

**Stage 4 Sampling: Within Household Selection**. A roster of each household will be collected and the relationships between all people age 14 and over will be obtained. A family will be randomly selected from the household where families with more generations are more likely to be selected. Then within the sampled family, 1 person from each generation present will be selected at random.

Generations are defined by the following age ranges:

- Teen: 14-17
- Young adult: 18-39
- Middle-aged adult: 40-59
- Older adult: 60+

### 5. Nonprobability-based design features to include recent immigrants

To adequately represent a diverse group of the largest and fastest growing immigrant populations in New Jersey, recruitment activities are focused on families with at least one first- or second-generation immigrant from China, Dominican Republic, Haiti, India, Jamaica, Korea, Mexico, Nigeria, or the Philippines. The inclusion criteria also include those who entered the US seeking

asylum, under temporary protected status, or related immigration authorities. Consequently, to enhance the recruitment of households with immigrants, the design has adapted respondent-driven sampling (RDS), a non-probabilistic, purposeful sampling technique that is used to recruit members of populations that cannot feasibly be recruited using probabilistic methods. RDS recruitment begins with "seeds," who are members of a focal community, to participate in the study. NJHealth Study seeds are recruited from the probability sample, when available, and in partnership with community-based organizations that are actively engaged with the groups of interest.

Initially, any New Jersey household with least one member who is a first- or second-generation immigrant is eligible for inclusion in the immigrant sample. The design then relies on two procedures to concentrate this sample on the specific immigrant groups. First, recruitment activities are conducted with community partner organizations. Second, immigrant sample study participants are asked to refer up to three additional families with immigrant members. They will be permitted to refer immigrant families from non-focal groups, but those participants will not be asked to provide further referrals. It is suggested, but not required, that they refer their own family members who live in New Jersey but not in their household (e.g., a parent or grandparent). The composition of the immigrant sample is monitored and recruitment strategies further adjusted (e.g., by varying the intensity of joint-recruitment activities with community partners) and inclusion criteria (e.g., by limiting eligibility to households with first generation immigrants) over time to ensure a balanced immigrant sample.

## 6. Summary

In summary, the New Jersey Population Health Cohort Study (NJHealth) is designed to improve the understanding and generate knowledge about factors that affect population health and opportunities to advance equity promoting policies in New Jersey and other populations. The study will collect survey responses, DNA, biomarkers, actigraphy and movement data, and link to other granular data over time on stress, resilience, trauma and health and well-being outcomes from a broad cross-section of the population across multiple generations, with additional targeting of low-income residents and diverse immigrant groups.

Study outcome domains include psychosocial components and measures of physical, cognitive, and behavioral health. Deep assessments of stress, anxiety, depressive symptoms, loneliness, and related factors are to be studied. Additional attributes under study include metabolic, clinical, sleep, cognition, physical performance and function, BMI, and medical comorbidities. Attention is also given to the use of essential health care services, and access/barriers encountered. Study research topics under consideration include: How do specific life events impact health and well-being?; What factors contribute to resiliency and positive health outcomes, despite or due to stressful experiences?; How do intergenerational family dynamics impact these relationships?; What roles do immigrant experiences play in exposure to stress, resilience, and population health outcomes?; and What roles do discrimination based on race and ethnicity play in population health and health equity?

The initial sample design objective was to select a representative sample of approximately 10,000 participants from broad sections of the population in the state. Adjustments to these sample targets were recently made to support the inclusion of an arm adapting RDS methods to enrich the sample with members of immigrant groups too rare to identify in the probability sample. Revised sample size targets currently consist of approximately 4,800 responding households and 6,000 completed survey interviews obtained from the probability-based sample selection process. The remaining respondents will be obtained from the purposefully recruited sample adapting RDS methods. Individuals sampled via these non-random selection schemes will limit their generalizability but have the potential to support behavioral analyses of understudied groups.

**Acknowledgements**

This effort reflects a collaboration between the Institute for Health, Health Care Policy, and Aging Research, Rutgers University and RTI International. Special acknowledgement goes to Stephanie Bergren, Margaret Koller, and other members of the NJHealth Study leadership team for their ongoing consultations, study management and study oversight. We are also grateful to the Robert Wood Johnson Foundation (Grant #7783) for providing funding for the design and implementation of the NJHealth Study. Additional support for NJHealth is provided by the State of New Jersey and Rutgers Health. The views expressed here do not necessarily reflect the views of the Foundation.

**References**

American Association for Public Opinion Research (2016). Address-based sampling. Report prepared for AAPOR Council by the Task Force on Address-based Sampling; R. Harter, Chair. Oakbrook Terrace, IL: AAPOR. https://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-(2).pdf

Cantor, J.C., Mouzon, D.M., Hu, W.T. et al, (2023). Health Implications of Enduring and Emerging Stressors: Study: Design of the New Jersey Population Health Cohort (NJHealth) Study. Draft in review.

Gile, K.J., Beaudry, I.S., Handcock, M.S., & Ott, M.Q. (2018). Methods for Inference from Respondent- Driven Sampling Data. Annual Review of Statistics and its Application, 5, 65-93.

Handcock, M.S (2017). Discussion of "Adaptive and network sampling for inference and interventions in changing populations" by Steve K. Thompson. Journal of Survey Statistics and Methodology, 5, 29–33.

Harter, R., & McMichael, J. P. (2012). Scope and coverage of landline and cell phone numbers appended to address frames. In Joint Statistical Meetings Proceedings, Survey Research Methods Section (pp. 3651–3665). Alexandria, VA: American Statistical Association.

Harter, R., McMichael, J., Brown, D., Amaya, A. E., Buskirk, T., & Malarek, D. (2018). Telephone appends for address-based samples—An introduction. (RTI Press Publication No. OP-0050- 1802). Research Triangle Park, NC: RTI Press. https://www.rti.org/rti-press-publication/telephone-appends-address-based-samples%E2%80%94-introduction

Heckathorn, D. D. (1997), "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," Social Problems, 44, 174–199.

McMichael, J. P., & Wiant, K. F. (2019). Improvements in sample design with address-level prediction models. American Association for Public Opinion Research Annual Conference, Toronto, Canada.

Salganik, M. J., and D. D. Heckathorn (2004). "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," Sociological Methodology, 34, 193–239.

Shook-Sa, B. E., Currivan, D. B., McMichael, J. P., & Iannacchione, V. G. (2013). Extending the coverage of address-based sampling frames: beyond the USPS Computerized Delivery Sequence File. Public Opinion Quarterly, 77, 994-1005.

Volz, E., and D. D. Heckathorn (2008), "Probability Based Estimation Theory for Respondent Driven Sampling," Journal of Official Statistics, 24, 79–97.

Wiant, K. F., & McMichael, J. P. (2019). Evaluating strategies for identifying rare populations within an address-based sample. Paper presented at American Association for Public Opinion Research Annual Conference, Toronto, Canada.

Zimmer, S., Morton, K and Cohen, S.B. (2022). New Jersey Population Health Cohort Study Sampling Design Plan. Project Report submitted to the Institute for Health, Health Care Policy, and Aging Research Rutgers University.