carelonresearch.com

Using individual-level race and ethnicity to improve representativeness in claims-based survey sampling for a survey of women with vasomotor symptoms of menopause

Judith J. Stephenson*, SM; Brian J. Harty, MA; Rebecca Conto; MA; Mia Si, PhD
Biruk Sisay, MS; Amanda Marshall, MBA; Sarah E. Daugherty, PHD
*Presenting author
Carelon Research, Wilmington, DE

PCR74

Background

- Identifying survey samples that are representative of the racial and ethnic distribution of a population of interest can be challenging using administrative claims data as the sampling frame because individual-level race and ethnicity data are often not available.
- Prior to 2023, individual-level race and ethnicity data were unavailable in Carelon Research's Healthcare Integrated Research Database (HIRD©), contributing to survey sample populations that were disproportionately White, non-Hispanic.
- With the 2023 addition of individual-level race and ethnicity data into the HIRD, we wanted to extract a more demographically aligned survey sample of women for a vasomotor symptom (VMS) study.

Objective

- To develop a sampling strategy algorithm, incorporating individual-level race and ethnicity, for creating a patient survey sample list from a source population that will result in a respondent population that aligns with the racial and ethnic distribution of a reference target population.

Methods and results

Data Source and Source Population:

- Data Source: The HIRD, a large administrative claims database, served as the sampling frame.
- Source Population: Survey-eligible women with commercial health insurance between the ages of 40 and 65 years at the beginning of the patient identification period (PIP) from 31 AUG 2023 through 30 AUG 2024 with no pregnancy-related or in-vitro fertilization (IVF) claims during the PIP.

Survey Sample Diversification Algorithm

STEP 1 Identify Reference Population.

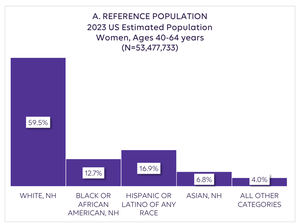
2023 US Census Estimate: Women, ages 40-64 years.
https://www2.census.gov/programs-surveys/popest/tables/2020-2023/national/asrh/nc-est2023-asr6h.xlsx

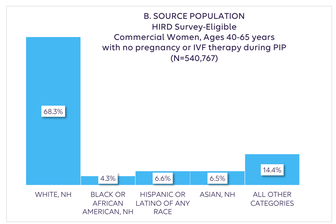
STEP 2 Align Reference and Source Populations' race & ethnicity categories; document any discrepancies.

Reference and Source Population race & ethnicity categories align; however, Reference Population age range is 40-64 years, and Source Population age range is 40-65 years.

STEP 3 Summarize percentage of population within each race & ethnicity category for both Reference and Source Populations; aggregate race & ethnicity categories with low frequencies (e.g., categories less than 5% of the sample).

Figure 1: Reference and Source Population Proportions by Race and Ethnicity Categories





Note: IVF=in-vitro fertilization; HIRD=Healthcare Integrated Research Database; NH=non-Hispanic; N=number; PIP=patient identification period between 31 Aug 2023 and 30 Aug 2024; US=United States

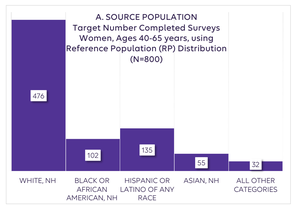
Due to very low frequencies, the American Indian or Alaska Native, NH; Native Hawaiian or Other Pacific Islander, NH; Other Race; and Unknown or Missing categories were combined into the category called 'All Other Categories'.

STEP 4 Apply Reference Population race & ethnicity distribution to Source Population's target number of completed surveys

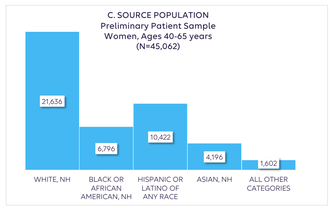
STEP 5 Apply Historic Race and Ethnicity List Completion Rates (LCRs) based on averages across 6 previously conducted survey studies to calculate size of preliminary patient survey sample required for patient recruitment and survey fielding.

Figure 2: Source Population Sample Before & After Application of Historic List Completion Rates (LCRs)

A. Source Population Target Number Completed Surveys With Reference Population Race and Ethnicity Distribution → B. Application of Historic LCR Rates → C. Source Population Preliminary Patient Sample



B. Historic Race & Ethnicity List Completion Rates (LCRs)	
White, NH	2.2%
Black or AA, NH	1.5%
Hispanic or Latino of any race	1.3%
Asian, NH	1.3%
All Other Categories	2.0%

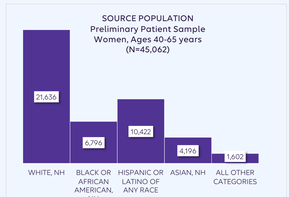


Methods and results continued

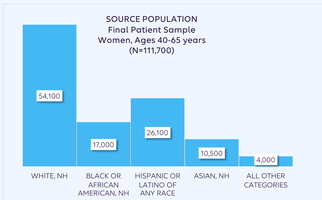
STEP 6 Apply quota oversampling rate to the Source Population PRELIMINARY Patient Sample to account for uncertainty in estimated percentage of women excluded because they reported not experiencing perimenopause or menopause signs or symptoms.

Figure 3: Source Population Patient Sample Before & After Application of Quota Oversampling Rate

A. Source Population PRELIMINARY Patient Sample Before Application of Quota Oversampling Rate → B. Application of Quota Oversampling Rate → C. Source Population FINAL Patient Sample After Application of Quota Oversampling Rate (rounded)



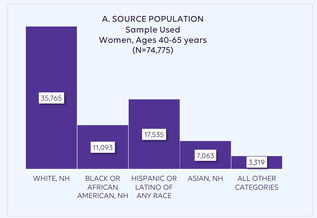
Quota Oversampling Rate	
White, NH	40.0%
Black or AA, NH	40.0%
Hispanic or Latino of any race	40.0%
Asian, NH	40.0%
All Other Categories	40.0%

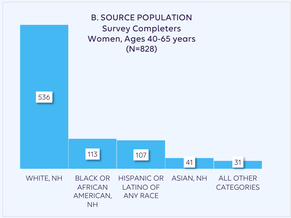


STEP 7 Apply stratified random sampling to FINAL Patient Sample List to create two subsamples consisting of a 70% Patient Sample to use for patient recruitment & survey fielding & a 30% Patient Sample to hold in reserve. Use the 70% Patient Sample for survey fielding. After survey fielding has ended, identify the number and percent of survey completers in each race & ethnicity category.

Figure 4: Source Population Sample Used for Survey Fielding and Survey Completers

A. Source Population 70% Sample Used for Survey Fielding → B. Source Population Survey Completers Using HIRD Race and Ethnicity → C. Source Population Survey Completers Race and Ethnicity Distribution





C. Survey Completers Race and Ethnicity Distribution (HIRD Race and Ethnicity) N=828	
White, NH	64.7%
Black or AA, NH	13.6%
Hispanic or Latino of any race	12.9%
Asian, NH	5.0%
All Other Categories	3.7%

STEP 8 Comparison of Current Study Patient Response to Historic Patient Response.

Figure 5A: Historic Patient Response

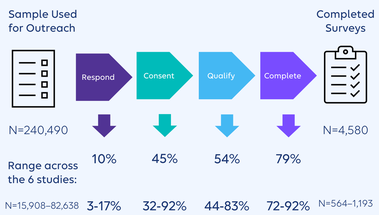
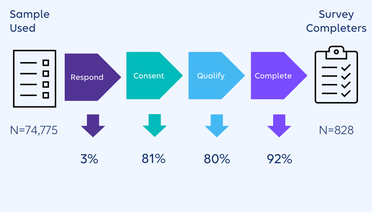


Figure 5B: Current Study Patient Response



STEP 9 Compare Current Study Race & Ethnicity Distribution to Reference, Source, and Historic Race & Ethnicity Distributions.

Table 1: Comparison of Current Study Race & Ethnicity Distribution to Reference, Source, and Historic Race & Ethnicity Distributions

	Reference Population Race and Ethnicity: 2023 US Women, Ages 40-64	Source Population Race and Ethnicity: HIRD Women, No Preg/IVF Ages 40-65	Source Population Historic Race and Ethnicity (6 surveys) All Ages	Source Population Current Study Race and Ethnicity: Women, Ages 40-65
Number of individuals	53,477,733	540,767	4,580	828
White, NH	60%	68%	86%	65%
Black or AA, NH	13%	4%	6%	14%
Hispanic or Latino of any race	17%	7%	2%	13%
Asian, NH	7%	7%	1%	5%
All other categories	4%	14%	5%	4%
Total	101%	100%	100%	101%

Note: AA = African American; NH = non-Hispanic; All other categories = American Indian/Alaska Native, NH; Native Hawaiian/Other Pacific Islander, NH; Other Race; Unknown/Undisclosed/Missing; Totals don't add to 100% because of rounding

- Compared with Carelon Research's historic patient response (Figure 5A), the current study response was lower (3% vs 10%), but a greater proportion of women consented (80% vs 54%), qualified (80% vs 65%) and completed a survey (92% vs 79%) (Figure 5B).
  - The response was lower because women responded so quickly the survey closed before many women had a chance to respond.
  - We hypothesize the consent, qualification, and completion rates are higher because women's health/menopause is a topic of interest to women.
- Compared with Carelon Research's prior survey respondent populations (86% White, non-Hispanic), the current survey achieved a markedly more diverse respondent population.
  - Of 828 completed surveys, 65% of women were classified as White, non-Hispanic, and 35% were classified as other racial and ethnic groups (Table 1).
  - Application of the sample diversification algorithm substantially improved alignment with the national Reference Population (59.5% White, Not Hispanic or Latino, and 40.5% other racial and ethnic groups).

Limitations

- The use of different Reference and Source Populations and/or Time Periods may give different results.
- The Source Population was limited to women with employer-sponsored commercial health insurance; the results can only be generalized to the US population that is covered by such insurance.

Conclusions

- The addition of individual-level race and ethnicity data to Carelon Research's large administrative claims database made possible the development of a sample diversification algorithm that incorporated historical survey metrics to generate a more representative and inclusive survey sample.
- Use of this more representative survey sample yielded a substantially more diverse respondent population than prior claims-only survey sample identification methods.

Funding & disclosures

Judith Stephenson, Principal Scientist
judith.stephenson@carelon.com

Poster presented at ISPOR 2026
May 17-20, Philadelphia, PA