carelon

carelonresearch.com

Assessing the representativeness of real-world claims databases

Judith J. Stephenson, SM; Chia-Chen Teng, MS; Katherine M. Harris, PhD
Carelon Research, Wilmington, DE, USA

RWD100

Background

- Healthcare claims databases are increasingly used to generate evidence for making healthcare decisions from real-world data (RWD).[1]
- However, little attention has been paid to the assessment of bias within the data and how it may affect results.
- There is greater potential for bias when non-representative RWD are used.[2]
- The question of how representative their data are is one that Carelon Research is often asked about their large claims database.
- The ability to assess the completeness and representativeness of RWD are important in obtaining generalizable results.[3]
- This study presents a method for assessing the representativeness of a real-world claims database versus a benchmark target population.

Objective

- To assess the representativeness of the Healthcare Integrated Research Database (HIRD®), a large, United States (US) healthcare database, using the 2020 US Census population as the benchmark target population.

Methods

Data sources

US Census data

- US Census data are collected by the US Census Bureau every 10 years about the US population.
- The data include counts or estimates of the total US population, including their sex, age, region, and race and ethnicity population distributions.
- The 2020 US Census total population is 331,449,281 people.

The HIRD

- The HIRD is a large healthcare database maintained by Carelon Research for health-related research.
- It includes claims data from 14 health plans in the Northeast, Midwest, South, and West geographic regions of the US, and is updated monthly.
- The HIRD researchable database contains administrative claims for Elevance Health members with commercial and Medicare health insurance.
- The 2020 HIRD researchable total population is 24,774,264 people or 7.5% of the 2020 US total population.

Assessment method

1. Identify a benchmark target population to be the comparison population for the HIRD.
2. Ensure the time periods are as comparable as possible.
3. Determine a set of characteristics for which the probability distributions exist for both populations.

Methods

Assessment method, continued

4. For each variable, calculate the difference between the probability distributions and the percent overlap of the two probability distributions.
5. Use the results of the comparisons to make statements about the representativeness of the HIRD population relative to the benchmark data population.

Outcome measures

Standardized mean difference (SMD)[4]

- The SMD assesses the magnitude of the difference or effect size between the probability distributions of common characteristics (e.g., sex, age) represented in both populations.
- The SMD is defined as:

$$SMD = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\frac{[\hat{P}_1(1-\hat{P}_1) + \hat{P}_2(1-\hat{P}_2)]}{2}}}$$

- Cohen's[5] interpretation of the magnitude of SMDs is used where:
  - ✓ 0.2 represents a small effect size
  - ✓ 0.5 represents a medium effect size
  - ✓ 0.8 represents a large effect size

Overlap Index (η)[6]

- The Overlap Index (η) quantifies the percentage overlap between characteristics defined by probability distributions in both populations.
- η is defined as:

$$\eta(A,B) = (1 - \frac{1}{2}\sum|fA(x) - fB(x)|) * 100$$

where fA(x) and fB(x) are two probability distributions.

η = 0% means the two distributions are completely separated, and η = 100% means the two distributions are completely the same.

We determined the representativeness of the HIRD researchable population compared to the US Census population.

- The 2020 HIRD researchable population was compared to the 2020 US Census population.
- The set of comparison characteristics were sex, age (5-year categories), geographic region, and race and ethnicity.
- SMD and η were calculated to determine the size of difference and percentage overlap between the census and HIRD variables.
- Statements about the representativeness of the HIRD researchable population compared to the US Census population were made based on the results of the comparisons.

Results

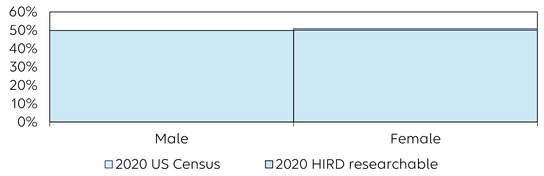
Comparison of 2020 US Census population and 2020 HIRD researchable population

Table 1. Sex: SMD = 0.02;

	2020 US Census	2020 HIRD researchable
Total population, N	331,449,281	24,774,264
Sex, %		
Male	49.1%	49.9%
Female	50.9%	50.1%

Note: Black font comparisons mean the HIRD Researchable percentage is greater than the US Census percentage; blue font comparisons mean the opposite.

Figure 1. Sex: Overlap = 99.2%



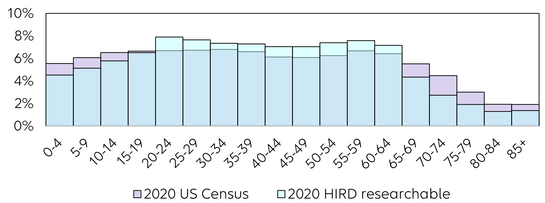
- The difference between the sex distributions of the census and HIRD populations is very small (**Table 1**), and the overlap is very large (**Figure 1**).
- The HIRD sex distribution is representative of the census sex distribution.

Table 2. Age (5-year categories): SMD = 0.19

	2020 US Census	2020 HIRD researchable
Total population, N	331,449,281	24,774,264
Age (5-yr cat), %		
Under 5 years	5.6%	4.5%
5 to 9 years	6.1%	5.1%
10 to 14 years	6.5%	5.8%
15 to 19 years	6.6%	6.5%
20 to 24 years	6.7%	7.9%
25 to 29 years	6.7%	7.7%
30 to 34 years	6.8%	7.3%
35 to 39 years	6.6%	7.3%
40 to 44 years	6.1%	7.1%
45 to 49 years	6.1%	7.0%
50 to 54 years	6.3%	7.4%
55 to 59 years	6.7%	7.6%
60 to 64 years	6.4%	7.2%
65 to 69 years	5.5%	4.3%
70 to 74 years	4.5%	2.7%
75 to 79 years	3.0%	1.9%
80 to 84 years	1.9%	1.3%
85+ years	1.9%	1.4%

Note: Black font comparisons mean the HIRD researchable percentage is greater than the US Census percentage; blue font comparisons mean the opposite.

Figure 2. Age (5-year categories): Overlap = 92.0%%



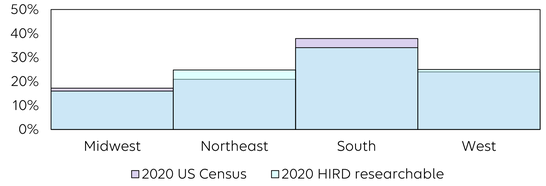
- The difference between the 5-year age category distributions of the census and HIRD populations is small (**Table 2**), and the overlap is large (**Figure 2**).
- The HIRD age distribution is representative of the census age distribution.

Table 3. Geographic Region: SMD = 0.16

	2020 US Census	2020 HIRD researchable
Total population, N	331,449,281	24,774,264
Geographic region, %		
Valid N	331,449,281	24,336,209*
Midwest	17.2%	16.0%
Northeast	20.9%	24.8%
South	37.9%	34.1%
West	24.0%	25.0%

*Excludes missing
Note: Black font comparisons mean the HIRD researchable percentage is greater than the US Census percentage; blue font comparisons mean the opposite.

Figure 3. Geographic Region: Overlap = 94.8%



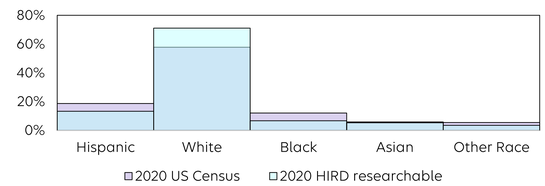
- The difference between the geographic region distributions of the census and HIRD populations is small (**Table 3**), and the overlap is large (**Figure 3**).
- The HIRD geographic region distribution is representative of the census geographic region distribution.

Table 4. Race and Ethnicity: SMD = 0.66

	2020 US Census	2020 HIRD researchable
Total population, N	331,449,281	24,774,264
Race and ethnicity, %		
Valid N	331,449,281	19,918,329*
Hispanic or Latino	18.7%	13.3%
White	57.8%	71.1%
Black or African American	12.1%	6.7%
Asian	5.9%	5.3%
Other Race	5.5%	3.5%

*Excludes missing
Note: Black font comparisons mean the HIRD researchable percentage is greater than the US Census percentage; blue font comparisons mean the opposite.

Figure 4. Race and Ethnicity: Overlap = 86.8%



- The difference between the race and ethnicity distributions of the census and HIRD populations is medium (**Table 4**), and overlap is medium large (**Figure 4**).
- The HIRD race and ethnicity distribution is somewhat representative of the census race and ethnicity distribution; the biggest difference is a more than 13% overrepresentation of white members in the HIRD researchable population.

Limitations

- Race and ethnicity was not determined the same way in the two populations. Census race and ethnicity was determined by self-report; HIRD race and ethnicity was obtained from a variety of sources and may not be directly comparable to the self-reported census race and ethnicity.
- Race and ethnicity data were missing for roughly 20% of HIRD members and may have limited the interpretability of the comparison; for example, if most or all the missing data were from non-white members.

Conclusions

- Overall, we found the 2020 HIRD researchable population to be representative of the 2020 US Census population in terms of sex, age, and region.
- However, race and ethnicity should be interpreted with caution due to differences in the way race and ethnicity was determined in the two populations and the large number of people missing race and ethnicity data in the HIRD.
- If variables are found to be not representative due to large SMDs and/or low overlap, post hoc weighting can be used to correct the imbalance
- This method provides a simple and practical empirical framework for further exploration of the representativeness of RWD sources and the generalizability of their results.

References

1. Dahlen A, Charu V. Analysis of sampling bias in large health care claims databases. *JAMA Network Open*. 2023; 6(1): e2249804. https://doi.org/10.1001/jamanetworkopen.2022.49804
2. Overbeek JA, Swart KMA, Houben E, Penning-van Beest, FJA, Herings RMC. Completeness and representativeness of the PHARMO general practitioner (GP) data: A comparison with national statistics. *Clin Epidemiol* 2023 Jan 5; 15:1-11. https://doi.org/10.2147/CLEP.S389598
3. Song F, Zang, C, Ma X, et al. The use of real-world data/evidence in regulatory submissions. *Contemp Clin Trials*. 2021; 109:106521. https://doi.org/10.1016/j.cct.2021.106521
4. Yang D, Dalton JE. A unified approach to measuring the effect size between two groups using SAS®, Paper 335-2012. Accessed August 10, 2023.
5. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd Ed. Lawrence Erlbaum Associates, 1988.
6. Pastore M, Calcagni A. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*. 2019; 10:1089. https://doi.org/10.3389/fpsyg.2019.0189

Funding and Disclosures

No external funding was received for the conduct of this study. All authors are employees of Carelon Research. JJS and KMH are shareholders of Elevance Health.

Contact information:
Judith Stephenson
judith.stephenson@carelon.com

Poster presented at ISPOR 2024, May 5-8, Atlanta, GA, USA.