carelonresearch.com

Representativeness of Linked Claims–EHR Data: Claims-Only vs Claims+EHR Populations

Lauren E Parlett, PhD; Amita Ketkar, SM, BDS; Judith J. Stephenson, SM; Michael Grabner*, PhD; Katherine M. Harris, PhD; Vincent J. Willey, PharmD

*Presenting author

Carelon Research, Wilmington, DE

RWD172

Background & Objectives

- HEOR studies often require the integration of administrative claims with electronic health records (EHR) to provide insights into patient characteristics and clinical and economic outcomes.
- As integrated data proliferates, questions should be raised about the underlying purpose of the integrations and any impact on patient characteristics in the resultant dataset.
- We compared characteristics from a claims-only sample of patients versus a subset of patients with both claims and EHR data.

Methods

Data Source

The Healthcare Integrated Research Database (HIRD®) is a large US claims & EHR database curated for health-related research. Details about the HIRD's covered population, data structure, data provenance and quality, and example applications have been previously published (Barron 2025).

Population

HIRD members were required to have continuous medical enrollment from January 1, 2024 to December 31, 2024. Database members with at least one 2024 EHR encounter were labeled "claims+EHR"; otherwise, they were "claims only".

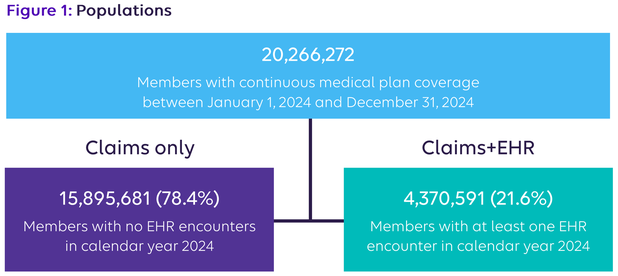
Variables

- Quan-Charlson comorbidities and the Quan-Charlson Index (QCI) are based on presence of ≥1 medical claim with the relevant ICD-10-CM codes in any position (Quan 2005) during 2024.
- Race/ethnicity is identified from multiple sources including self-report and imputation (Price 2025).
- Area-level socioeconomic status (SES), derived from the 2018-2022 5-year estimates of the American Community Survey, was linked by database member residence at the census-block group level.

Analysis

- For continuous variables, mean, standard deviation, median, and interquartile range are reported. To assess difference across populations, the standardized mean difference was computed (Yang 2012).
- For categorical measures, frequency and proportion are reported. The population probability distributions were compared using the overlap index (ŋ) where 0% means no overlap and 100% means complete overlap (Pastore 2019).

Figure 1: Populations



Conclusion

Large, integrated claims plus EHR datasets provide opportunities to address clinically-focused evidence needs. Analyses assessing any differences in individuals' characteristics in the integrated dataset compared to the source population are critical when interpreting and applying study findings.

Table 1: Sociodemographic characteristics

Characteristic	Claims Only		Claims + EHR		SMD or ŋ
	N	%	N	%	
Age, years					84.4
0-17	3,216,526	20.2	630,151	14.4	
18-44	6,273,598	39.5	1,297,489	29.7	
45-64	4,611,461	29.0	1,449,017	33.2	
65+	1,793,837	11.3	993,906	22.7	
Mean, SD	38.2	90.1	45.9	26.0	0.26
Median, IQR	38	21-55	49	29-63	
Sex					91.0
Male	7,703,892	48.5	2,513,947	57.5	
Female	8,183,211	51.5	1,856,101	42.5	
Missing	8,578	0.1	543	0.0	
Race					85.4
NH American Indian or Alaska Native	55,842	0.4	12,748	0.3	
NH Asian	1,204,066	7.6	164,370	3.8	
NH Black or African American	1,319,551	8.3	387,159	8.9	
Hispanic or Latino of any race	2,065,077	13.0	332,994	7.6	
NH Native Hawaiian or Other Pacific Islander	30,515	0.2	5,439	0.1	
NH Other race	470,259	3.0	81,794	1.9	
NH White	10,056,461	63.3	3,379,346	77.3	
Unknown or Undisclosed	693,910	4.4	6,741	0.2	
Census Region of member residence					77.4
Midwest	3,109,144	19.6	1,845,505	42.2	
Northeast	2,858,153	18.0	643,307	14.7	
South	5,264,900	33.1	1,244,023	28.5	
West	4,654,793	29.3	636,957	14.6	
Missing	8,691	0.1	799	0.0	
Insurance Coverage					87.2
Commercial	14,773,804	92.9	3,500,035	80.1	
Managed Medicare / Supplemental	1,119,195	7.0	870,012	19.9	
Other	2,682	0.0	544	0.0	
AHRQ SES Index Quartile					95.6
1 (lowest quartile)	2,077,789	13.1	626,536	14.3	
2	3,155,312	19.9	978,934	22.4	
3	4,078,615	25.7	1,134,867	26.0	
4 (highest quartile)	5,435,464	34.2	1,296,423	29.7	
Missing	1,148,501	7.2	333,831	7.6	

ŋ (eta) = overlap index representing percentage of distribution overlap; AHRQ = Agency for Healthcare Research and Quality; IQR = interquartile range; NH = Non-Hispanic or non-Latino; SD = standard deviation; SES = socioeconomic status; SMD = standardized mean difference

Table 2: Comorbidity burden, healthcare costs, and coverage utilization

Characteristic	Claims only		Claims + EHR		SMD	ŋ
	Mean/N	SD/%	Mean/N	SD/%		
QCI in 2024						
Mean, SD	0.3	0.9	0.7	1.4		0.39
All-cause total medical and Rx costs (USD)						
Mean, SD	5,755	28,359	11,400	36,888	0.19	NA
Median, IQR	887	164 -3,353	2,757	899 -9,038		
Healthcare coverage utilization						
1+ medical claim, N, %	12,863,054	80.9	4,297,468	98.3	91.3	
Medical claim count	11.2	18.7	20.5	26.1		0.45
1+ Rx claim, N, %	7,771,908	48.9	3,001,702	68.7	90.1	
Rx claim count	5.1	9.7	10.8	14.9		0.52

ŋ (eta) = overlap index; USD – 2024 United States dollars; QCI = Quan Charlson Index; Rx = prescription

Figure 2: Prevalence of Quan Charlson Comorbidities, %

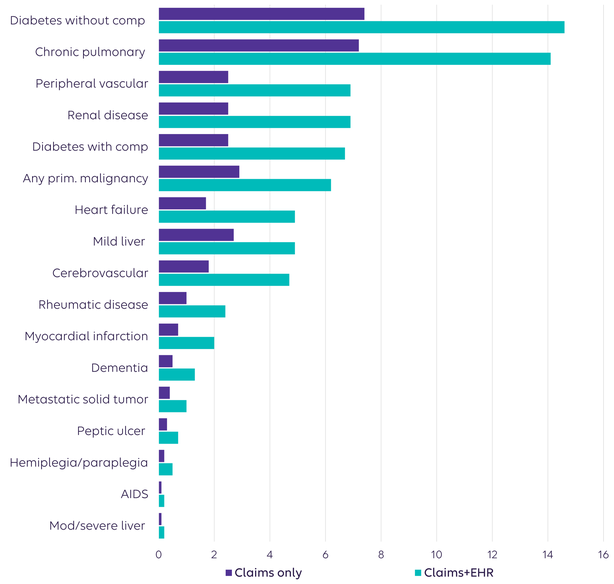
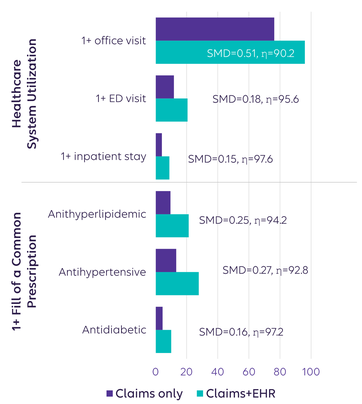


Figure 3: Quan Charlson Index Categories, % of population

Figure 4: Healthcare System Utilization and Prescription Fills. %



ŋ (eta) = overlap index; ED = emergency department; SMD = standardized mean difference

Table 3: Selected biomarkers

	Claims+EHR with at least one record		Among those with at least 1 record, distinct* record count in 2024	
	N	%	mean	SD
Weight	2,809,753	64.3	2.7	2.8
Height	2,622,554	60.0	2.6	2.6
Body mass index	2,554,916	58.5	2.5	2.4
Blood pressure	2,520,497	57.7	5.1	5.3

SD – standard deviation; * distinct based on encounter date

Results

- Of the over 20 million HIRD members with continuous medical coverage in 2024, one in five also had at least one EHR encounter (Figure 1).
- The largest sociodemographic differences were Census region of member residence, race/ethnicity, and insurance coverage type (Table 1). Members with claims+EHR tended to be more likely non-Hispanic White, less likely to be Hispanic or Latino, more likely enrolled in a managed Medicare plan, and much more likely to reside in the Midwest than members with claims-only data.
- Across all Quan-Charlson comorbidities (Figure 2), the claims+EHR population had higher 2024 prevalences than the claims-only population. Mild liver disease prevalence (claims+EHR: 4.9%; claims-only: 2.7%) was closest with a 1.8x difference. Heart failure (claims+EHR: 4.9%; claims-only: 1.7%) had the widest relative gap of 2.8x. About 9.2% of the claims+EHR members had three or more QCI comorbidities in 2024 compared to 3.4% of the claims-only members (Figure 3).
- Members with claims+EHR had a higher proportion of the population with at least one medical claim compared to members with claims only (claims+EHR: 98%; claims-only: 81%). Similarly for Rx claims, claims+EHR members had a higher proportion with at least 1 Rx claim (claims+EHR: 69%; claims-only: 49%) (Table 2). .
- Across some common prescriptions, members with claims+EHR were about twice as likely to have filled them in 2024 compared to members with claims-only (Figure 4). Compared to members with claims-only data, members with claims+EHR also showed higher healthcare utilization for inpatient, ER, and office visits, and higher 2024 total costs (median: $2,757 versus $877) (Figure 4; Table 2).
- Recent BMI (59%), weight (64%), height (60%), and blood pressure (58%) measurements were available for most members with claims+EHR (Table 3) with many having at least two measurements during 2024. When blood pressure data were available, members with claims+EHR records had, on average, at least five separate encounter dates (SD: 5.3) with blood pressure readings.

Limitations

- Availability of EHR data for a given member of the HIRD is subject to several selection processes; the claims+EHR subgroup is not randomly determined. These selection processes may drive some of the differences in population characteristics.
- Results are derived for 2024; population characteristics may differ for prior and/or subsequent years.

References

1. Barron, J.J., Willey, V.J., Doherty, B.T., Tunceli, O., Waltz, C.R., Grabner, M., Beachler, D.C., Lanes, S. and Cziraky, M.J. (2025), The Healthcare Integrated Research Database (HIRD) as a Real-World Data Source for Pharmacoepidemiologic Research. Pharmacoepidemiol Drug Saf, 34: e70110. https://doi.org/10.1002/pds.70110
2. Pastore M, Calcagnì A (2019): Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. Front. Psychol. 10:1089. doi: 10.3389/fpsyg.2019.01089
3. Price A, Chi W, Overhage JM. Implementation and Validation of a Prioritization Logic to Identify the Best Available Race/Ethnicity Information for Members in Commercial Plans. Poster presentation at the 2024 ISPOR Annual Meeting in Atlanta, GA. Available at https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3898/139634 [Accessed 19 Mar 2026]
4. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005 Nov;43(11):1130-9. doi: 10.1097/01.mlr.0000182534.19832.83.
5. Yang D, Dalton J. A unified approach to measuring the effect size between two groups using SAS. SAS Global Forum 2012, Paper 335-2012. https://support.sas.com/resources/papers/proceedings12/335-2012.pdf [Accessed 19 March 2026]

Funding & disclosures

This work was supported by Carelon Research. All co-authors are employees of Carelon Research.

Michael Grabner, PhD
michael.grabner@carelon.com

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